Logic, Epistemology, and the Unity of Science 38

Juan Redmond
Olga Pombo Martins
Ángel Nepomuceno Fernández *Editors*

Epistemology, Knowledge and the Impact of Interaction

Springer

Logic, Epistemology, and the Unity of Science

Volume 38

Logic, Epistemology, and the Unity of Science aims to reconsider the question of the unity of science in light of recent developments in logic. At present, no single logical, semantical or methodological framework dominates the philosophy of science. However, the editors of this series believe that formal techniques like, for example, independence friendly logic, dialogical logics, multimodal logics, game theoretic semantics and linear logics, have the potential to cast new light on basic issues in the discussion of the unity of science.

This series provides a venue where philosophers and logicians can apply specific technical insights to fundamental philosophical problems. While the series is open to a wide variety of perspectives, including the study and analysis of argumentation and the critical discussion of the relationship between logic and the philosophy of science, the aim is to provide an integrated picture of the scientific enterprise in all its diversity.

More information about this series at http://www.springer.com/series/6936

Juan Redmond • Olga Pombo Martins
Ángel Nepomuceno Fernández
Editors

Epistemology, Knowledge and the Impact of Interaction

Springer

Editors
Juan Redmond
Instituto de Filosofia
Universidad de Valparaíso
Valparaíso, Chile

Ángel Nepomuceno Fernández
Department of Logic and Philosophy
University of Sevilla
Sevilla, Spain

Olga Pombo Martins
University of Lisbon
Lisboa, Portugal

ISSN 2214-9775 ISSN 2214-9783 (electronic)
Logic, Epistemology, and the Unity of Science
ISBN 978-3-319-26504-9 ISBN 978-3-319-26506-3 (eBook)
DOI 10.1007/978-3-319-26506-3

Library of Congress Control Number: 2016935705

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG Switzerland

Preface

As was pointed out by Shahid Rahman and John Symons in the preface to the first volume that launched in 2004 the now so successful series *Logic, Epistemology and the Unity of Science*, the notion of science incepted by modern and contemporary encyclopaedists emerged from the idea that logical and epistemological reflections on science should be informed by reflection on the dynamical and cooperative nature of science and scientific reasoning. The interactive and/or interdisciplinary feature of science provides one of the basic tenets underlying the papers of our volume. Indeed, the present volume contains a selection of articles that stress and delve into the complex interplay of scientific knowledge, interaction and reasoning by covering various areas of research such as epistemology: logic, argumentation theory, linguistics and philosophy of science. Accordingly, the contributions have not been grouped by disciplines but by six themes that constitute a whole triggered by diverse forms of cross-fertilization. Indeed the distribution could have been carried out in a different manner; however this has been our choice, may the reader choose his own way to establish the conceptual links that animate his own scientific motivations.

We shall present in the next section a brief overview of the contents of the six parts, but let us, before this, point out that the parts contain complementary perspectives sometimes even antagonistic ones. This, so to say, dialogical structure of the book yields the following organization

- While the first part delves into the problems of characterizing the epistemic features of inference and interaction, the second contains contributions based on model-theoretical approaches to the interface between knowledge, modal logic and interaction.
- The third part develops further the interactive perspectives of the first two chapters by studying the consequences of the deployment of an argumentative frame for interaction and by tackling the notion of meaning in relation to assertion-conditions in a context.
- The fourth part presents both an historical and critical interlude on the epistemological foundations underlying the precedent chapters. This chapter that discusses

Leonard Nelson's criticism could be read as complementary to the non-antirealist but nevertheless proof-theoretical approach of *Ludics* in Chapter 1. Moreover, the distinction between the justification of *inferred truths* and *direct evidence* might be also related to the differentiation between demonstration and proof-object essential to Constructive Type Theory as discussed in the first chapter.

- The last two parts discuss two crucial aspects of the links between scientific knowledge and epistemology, namely, the naturalization of epistemology and the nature and role of models from the perspective of sciences themselves.

Perhaps one way to see the organization principle behind the different chapters is to understand them as deploying the interplay between different aspects of the logic of knowledge and scientific (interdisciplinary) reasoning, within the frame of a suitable theory of meaning and interaction.

The Dynamics of Knowledge I: Proof-Theoretical Approaches and the Interactive Viewpoint

Chapter 1 starts with the paper *Perennial Intuitionism*, penned by Johan G. Granström, who claims that the old antagonism between realism, conceptualism and nominalism, nowadays reflected in the tension between model theory, intuitionism and formalism, can be solved by Per Martin-Löf's *Intuitionistic Type Theory*. The leading idea is that intuitionistic type theory can be understood as a kind of moderate realism in the style of the Aristotelian-Thomist tradition.

Chapter 2 The paper of Thomas Piecha and Peter Schroeder-Heister delves into the epistemic roots of proof theory by developing one of the most difficult issues for proof-theoretical approaches, namely the question about the proof-theoretical interpretation of atomic propositions. More precisely, the authors of *Atomic Systems in Proof-Theoretic Semantics: Two Approaches* compare two different approaches to atomic systems in proof-theoretic semantics: one that is compatible with an interpretation of such systems as representations of states of knowledge, and another that takes atomic systems to be definitions of atomic formulas. Both approaches lead to different concepts of proof-theoretic validity: a consequence relation that is monotone (with respect to extensions of atomic systems) or a relation that does not verify transitivity and is not monotonic.

Chapter 3 Shahid Rahman, Radmila Jovanovic and Nicolas Clerbout discuss in their paper *Knowledge and its Game-Theoretical Foundations: The Challenge of the Dialogical Approach to Constructive Type Theory* how a dialogical approach to Martin-Löf's *Constructive Type Theory* (CTT) provides a fruitful and natural way of linking CTT with interaction. The authors motivate their study by two main case-studies where the notion of inter-dependence of quantifiers plays a crucial role, namely, the game-theoretical interpretation of the axiom of choice and of anaphora.

Chapter 4 Darryl McAdams and Jonathan Sterling contribute with a paper with the title *Dependent Types for Pragmatics*, where they propose solutions to

problems in pragmatics, such as pronominal reference and presupposition, based on Martin-Löf's Type Theory. In fact, the paper can be seen as pushing forward the dialogical formulation of the CTT approach to anaphora and pronouns discussed by Rahman/Jovanovic/Clerbout in the preceding chapter. Indeed, the authors introduce an operator called *require*, that has a clear interactive reading, and that they motivate not only by comparing it with the standard methods of Discourse Representation Theory and Dynamic Semantics, but also with the meaning explanations of CTT.

Chapter 5 The paper *On the Computational Meaning of Axioms* by Alberto Naibo, Mattia Petrolo and Thomas Seiller retakes another crucial issue to the proof-theoretical framework: the meaning of the axioms. Their approach is related to Ludics, developed by Jean-Yves Girard and collaborators, and goes deeper into the roots of meaning as interaction. The result is nevertheless neither typed nor necessarily anti-realist. In particular, unlike other proof-theoretical approaches, the standard framework of classical logic is not called into question.

The Dynamics of Knowledge II: Epistemology, Games and Dynamic Epistemic Logic

Chapter 6 In the paper *A Dynamic Analysis of Interactive Rationality* Eric Pacuit and Olivier Roy propose a general framework for the study of informational contexts. In fact the point is to elucidate how such contexts may arise in order to provide a comprehensive understanding of strategic interaction. According to their proposal, informational contexts are viewed as the fixed-points of iterated "rational responses" to incoming information about the agent's possible choices. Furthermore, the authors generalize existing rules for information updates used in the dynamic-epistemic logic literature, and this strategy is also applied to understand the notion of admissibility in general and to solve a well-known paradox of admissibility.

Chapter 7 Peter Hawke in *Relevant Alternatives in Epistemology and Logic* provides a survey of the diverse array of "relevant alternatives" theories of knowledge. The paper provides a schema in order to classify theories at different levels of abstraction and presents a sample of relevant alternatives theories by contrasting *question-first* and *topic-first* theories. The framework blends with current discussions in the philosophical literature and allows at the same time the study of different ways of formalizing some of the most important positions in the corresponding debates.

Chapter 8 Chenwei Shi in *Knowledge Based on Reliable Evidence* proposes to model a piece of evidence as a set of hypotheses supported by that piece of evidence. In other words, a set of hypotheses is seen as constituting a piece of evidence itself, if it is supported by a piece of evidence intrinsically linked to that set. This leads the author to develop an alternative version of reliability and to a novel approach to the notorious discussion of knowledge as justified belief.

More generally, by a systematic comparison between different kinds of beliefs, it is claimed that it is the proposed notion of reliability, instead of robustness, that qualifies belief as knowledge. Finally, the author explores the agent's knowledge updating, particularly triggered by evidence dynamics, and develops a suitable complete dynamic logic.

Chapter 9 Can Başkent in *Public Announcements and Inconsistencies: For a Paraconsistent Topological Model* develops public announcements logic (PAL) in topological contexts. Topology has been revealed to be a fruitful frame for the interpretation of logical systems such as the intuitionistic interpretation of a S4 modal system. This article gives another twist: PAL is studied within a paraconsistent topological model. The paper concludes by suggesting that some possible fruitful extensions of the proposal include the study of paradoxical public announcements (such as Moore-sentences) and mereology.

Chapter 10 Manuel Rebuschi in *Knowing Necessary Truths* deals with the problem of establishing the difference between knowing a necessary proposition and knowing that it is a necessary truth, a difference that constitutes the core of the so-called modal omniscience. The author considers that the standard two-dimensional semantics does not offer an adequate solution and proposes instead to make use of a modified version of Hintikka's notion of world lines. In fact, the proposed framework combines metaphysical possibilities à la Kripke with epistemically possible worlds à la Hintikka. In a way, it is a two-dimensional framework after all, but not a standard one.

Chapter 11 Emilio Gómez-Caminero and Angel Nepomuceno in *Modified Tableaux for Some Kinds of Multimodal Logics* develop semantic tableaux for multimodal logic. In order to treat the variety of accessibility relations the authors introduce *inheritance rules*, which can be adapted to important logical systems. Each logical system, containing the same type of modal operators, will have tableaux with different *inheritance rules*. The paper also discusses the use of a special kind of tableaux that allow the treatment of infinitary operators. The last two sections of the paper deal with more general issues on epistemology as a scientific discipline (naturalized epistemology) and on epistemological considerations on science. In fact, these last sections are linked to Parts **V** and **VI** of the present volume.

Argumentation, Conversation and Meaning in Context

Chapter 12 Silvia Martínez Fabregat's *Irony as a Visual argument* discusses the persuasive strength of irony in relation to its dependence upon the active interaction with a targeted audience. The author's approach on the uses of rhetorical tropes allude to the different ways that speakers should present their arguments depending on the argumentation field where they are working, the potential audience that they imagine or their argumentative goals. According to this view, the selection of a rhetoric strategy instead of any other defines the speaker as well as her argumentation. Furthermore, the author shows how irony operates into the written

speech by using, as an example, Joan Fuster's aphorism ("I do not understand who said that they underestimated money. It takes so much hard work to earn it!") and explores the possibilities of ironic argumentation in the visual field by focusing on one of Banksy's paintings.

Chapter 13 Sruthi Rothenfluch in *Ascribing Knowledge to Experts: A Virtue-Contextualist Approach* argues that traditional forms of contextualism, which employ relevant alternatives and sensitivity models, cannot accommodate our knowledge judgments in contexts of expert advice. She argues that contextualists must incorporate a virtue responsibilist approach to account for knowledge attributions and denials in such scenarios. What matters in such cases, according to Rothenfluch, are the subject's mastery of underlying principles of the field and appropriate application of such understanding. These features, while distinctive of virtue possession and of expert knowledge, cannot be measured by assessing a subject's epistemic response to counter-factual situations, and for this reason, is not captured by traditional models of contextualism.

Chapter 14 Gildas Nzokou in *Defeasible Argumentation in African Oral Traditions. A Special Case of Dealing with the Non-monotonic Inference in a Dialogical Framework* works out the thought-provoking structural correspondence between some specific oral legal debates of the African traditions and the non-monotonic reconstructions of (western) legal reasoning. The author's elegant development is based on providing a dialogical frame in which oral debates of the African tradition and the debates of western legal processes can be compared and studied. The paper includes some interesting brief remarks on the links between this kind of reasoning and the dialectical arguments as understood by Aristotle.

Chapter 15 Vít Punčochář in *Semantics of Assertibility and Deniability* reacts to Christopher Gauker's book *Conditionals in Context*. In his book, Gauker proposes to formulate a semantics based on the concept of "assertibility in a context" instead of "truth in a world". Primitive contexts are consistent sets of literals. Multicontexts of some level are sets which contain primitive contexts and/or multicontexts of lower levels. Though the author agrees in principle with the proposal he is less convinced by the non-compositionality of the underlying theory of meaning and by, as he sees it, lack of unity of meaning of the connectives that result of Gauker's framework. This leads Punčochář to follow a double strategy, firstly he reformulates Gauker's notion of context with the help of the notion of context of Robert Stalnaker, and then he develops a compositional semantics based on pairs of connectives, one extensional and the other intensional. This two-folded strategy preserves, according to the author of the paper, the unity of meaning of the logical connectives.

Chapter 16 In the last contribution of the chapter *The Quest for the Concept in the Twentieth Century: Predicates, Functions, Categories and Argument Structure*, Francisco J. Salguero-Lamillar points out that from the times of Ancient Greece to the most contemporary studies, researchers have attempted to decipher the mechanisms according to which concepts are defined from the meaning of words in such a way that these concepts and their form of apprehension reflect, respectively, the underlying ontology and epistemology deployed by these concepts. The main aim of the contribution of Salguero-Lamillar is to explore the seminal ideas that have resulted in categorial grammars and their relationship with other grammatical

models and actual theories of meaning, in a historical process that takes us from the notion of category to that of predication, and from this to the notion of function, then to functional categories and finally to the linguistic notion of argument structure.

A Critical Interlude

Chapter 17 is constituted by only one paper by Jan Woleński (*On Leonard Nelson's Criticism of Epistemology*), which underlies the importance of Leonard Nelson (1882–1927) in the history of contemporary philosophy, and his decisive role in the establishment of Neo-Kantism in the so-called New-Friesian School and the Badenian School. As mentioned in the first section of the preface, the main aim of Wolenski's paper is to present and rebuild both of Nelson's proofs against the possibility of epistemology. The author shows that both of Nelson's proofs share the following premise: *The fundamental task of epistemology consists in demonstrating objective truth or validity of human knowledge*. Wolenski points out that Nelson's arguments concern the impossibility of epistemology, but they do not say that knowledge cannot be achieved at all. Roughly speaking, Nelson argues that if we restrict knowledge to something indirect and obtainable by proof, that is, by assuming that every knowledge is inferred from another knowledge, we will inevitably fall into a dilemma. An appeal to direct perceptual knowledge gives no way out either, because it does not solve the question of justification of propositions. This reasoning suggests that the actual possibility of knowledge strongly depends on direct non-evident knowledge. Nelson's essential step rejects the identification of such a kind of knowledge with propositions.

The reader will be tempted to compare some of the insights contained in Nelson's criticism against the possibility of epistemology on one hand with Granström's defence of moderate realism (one might perhaps relate Nelson's remarks with the distinction between *proof-objects* – that provides the ontology that furnishes the truth of a proposition – as corresponding to Nelson's direct knowledge and *demonstrations* constituted by inferences between propositions) and on the other with the non-antirealist approach of Ludics and Geometry of Interaction discussed by Naibo, Petrolo and Seiller.

Knowledge and Sciences I: Naturalized Logic and Epistemology, Cognition and Abduction

Chapter 18 John Woods in *Logic Naturalized* provides a critical perspective on the possible marriage of logic and epistemology. Indeed, Woods points out that the mathematical turns in logic of the nineteenth century left out the human reasoner. Since then indifference to the realities of human cognitive agency is still retained. According to our author here is a clearly discernible pattern. The greater

the theory's interest in approximating to how humans actually think, the more complex is the theory's formal mechanisms. On this view, realist approximation varies proportionally with mathematical enrichment. A contrary view is suggested in this chapter inspired by Quine's proposal of a naturalized epistemology: the bold main proposal of Wood's paper is to bring forward a naturalization of logic *that might hold at least some of the promise that now graces philosophical work on knowledge*.

Chapter 19 The paper *Action Models for the Extended Mind* by Fernando Soler-Toscano can be seen as a reply or even as an acknowledgment of John Woods' plea for studies in logic that takes seriously into account the cognitive realities of human reasoners. Indeed, Soler-Toscano's paper proposes to pay attention to the relevance of the environment in the performance of cognitive tasks as shown by recent studies in cognition. The idea of the *extended mind* focuses on the importance of external resources that can be considered as part of the mind, and he proposes to make use of the instruments of dynamic epistemic logic in order to develop a logical system that is sensitive to cognitive-bounds of human agents. In fact, according to the author, a logical analysis of the epistemic actions related with the cognitive configuration and exploitation of the environment throws light on the novelties of the role of external resources.

Chapter 20 Valeriano Iranzo (*Explanatory Reasoning: A Probabilistic Interpretation*) analyzes abduction and inference to the best explanation (IBE) as forms of reasoning in the scientific research. This paper too can be seen as linked to Woods' criticism to the highly abstract modelizations of human reasoning. However, the subject here is not logic in itself but reasoning as deployed by abduction understood as IBE. In fact, the author distinguishes between discerning (a) which explanation is the best one and (b) whether the best explanation deserves to be legitimately believed. After discussing and contesting the reduction of uncertainty to definitions of explanatory power, the author proposes a rule, called "rule R1*", as a sufficient condition to discern which explanation is the best. In relation to (b), Iranzo proposes a probabilistic threshold as a minimal condition for entitlement to believe. The rule R1* and the threshold condition are intended as a partial explication of explanatory value (and, consequently, also as a partial explication of "inference to the best explanation").

Chapter 21 The chapter closes with the paper *The Iconic Moment. Towards a Peircean Theory of Diagrammatic Imagination* by Ahti-Veikko Pietarinen and Francesco Belluci, who stress the relevance of Peirce's understanding of imagination, abductive reasoning and diagrammatic representations for understanding crucial aspects of scientific reasoning and discovery. More precisely, as pointed out by the authors, in 1908 Peirce stated that deduction consists of "two substages", logical analysis and mathematical reasoning. Mathematical reasoning is again divisible into "corollarial and theorematic reasoning", the latter concerning an invention of a new icon, or "imaginary object diagram", while the former results from "previous logical analyses and mathematically reasoned conclusions". The iconic moment is clearly stated here, as well as the imaginative character of theorematic reasoning. But translating propositions into a suitable diagrammatic

language is also needed. Imagination becomes a crucial part of the method for attaining truth, that is, of the logic of science and scientific inquiry, so much so that Peirce took it that "next after the passion to learn there is no quality so indispensable to the successful prosecution of science as imagination". In this paper, the author investigates the aspects of scientific reasoning and discovery that seem irreplaceably dependent on a Peircean understanding of imagination, abductive reasoning and diagrammatic representations.

Knowledge and Sciences II: The Role of Models and the Use of Fictions

Chapter 22 The paper that launches the present chapter can be seen as the result of practicing a naturalization of philosophy of sciences. More precisely it is about how to "naturalize" the notion of *emergence*, that is subject of many researches in philosophy of science. As a matter of fact, in his paper *Does Emergence Also Belong to the Scientific Image? Elements of an Alternative Theoretical Framework Towards an Objective Notion of Emergence*, Philippe Huneman stresses on one hand the importance of the notion of emergence for the organization of our knowledge and on the other the many objections that have been raised against the coherence of this notion. In order to go out of this impasse, Huneman proposes to reformulate the notion of emergence by delving into its scientific roots. Furthermore, the author proposes to make use of a notion of *computational emergence* that can in addition be characterized in terms of causation. After having won this new notion of emergence, the author turns to the question of testing if scientific data support or not the existence of instantiations of such a concept.

Chapter 23 In his paper *A Comparison of the Semantics of Natural Kind Terms and Artifactual Terms*, Luis Fernández Moreno tackles one issue that is subject of a host of works in philosophy of science and beyond, namely the notion of artifact. Moreover, one thorny task in this context is to characterize the meaning of artifactual terms. The author not only examines the links between natural kind terms and artifactual terms but he also proposes a theory of meaning for the latter. To that end, the author discusses Hilary Putnam's semantics for terms of natural kinds and its extension to artifactual terms. Though Fernández Moreno agrees with this extension, the reference fixing theory he advocates differs from that of Putnam's. Furthermore, the author proposes a view on the meaning of artifactual terms, which conflicts with the one it would follow from extending to such terms Putnam's view. In fact, the semantic theory advocated by Fernández Moreno with respect to artifactual terms is one of the versions of the "traditional theory": the cluster theory.

Chapter 24 In his paper *Models, Representation and Incompatibility. A Contribution to the Epistemological Debate on the Philosophy of Physics*, Andrés Rivadulla tackles the issue of models in science from a non-realist perspective.

The author starts his paper by stressing the fundamental role that models play in nowadays science in general and in the methodology of theoretical physics in particular. According to him, there is no branch in contemporary physics, whether it be cosmology, astrophysics or microphysics, where such models are not used. These models are idealized constructs of a single phenomenon or about a limited empirical domain. They are intended to both save the phenomena and to make testable predictions about the domain they are concerned with. The main anti-realist manifesto of Rivadulla is that models are not susceptible to being true or verisimilar representations of certain aspects of reality. According to our author, models make use of extant theories and are of particular use in domains lacking theories. Moreover, in a historical sequence of theoretical models about a certain domain, not every model is compatible with previous ones. This is the case of Ptolemaic and Copernican cosmological models or of Einsteinian and Newtonian gravitational models. The incompatibility among models (and even theories) about the same domain is the most serious issue facing standard convergent realism. In order to illustrate and bring forwards arguments for his claim, the author focuses on various kinds of theoretical models employed by nuclear physics.

Chapter 25 In the last paper of the book (*Fictions in Legal Science: The Strange Case of the Basic Norm)*, Juliele Maria Sievers tackles the notion of fiction in the context of legal science. Her main strategy is similar to the one that Huneman applied to the notion of *emergence*. She studies the notion of fiction in legal science not as an instance of the philosophical notion of fiction but from the legal perspective and practice. More precisely, the main aim of the author is to analyze the use of fiction by the legal science under the light of the legal theory proposed by Hans Kelsen (1881–1973), especially concerning his proposal that the legitimization of the whole positive legal system is based on a fiction, called the Basic Norm (*Grundnorm*). The difference, according to Sievers, is that this "norm" must be seen as a methodological or scientific tool, and not as an ordinary norm among others in the legal system. Furthermore, her aim is to elucidate how such a notion of fiction can display that important normative function and still preserve the "principle of purity" of the Kelsenian legal theory.

Valparaíso, Chile — Juan Redmond
Lisboa, Portugal — Olga Pombo Martins
Sevilla, Spain — Ángel Nepomuceno Fernández

Acknowledgements

The present work is a result of the project Fondecyt Regular No. 1141260 (Principal Investigator: Juan Redmond). We thank the CONICYT (Chile) for fostering the preparation of this new LEUS volume.

Some of the papers developed out of presentations at the *International Symposium Epistemology, Logic and Language* are organized and fostered by the Center for Philosophy of Sciences of the University of Lisbon, which took place in Lisbon from 29 to 31 October 2012. The colloquium was put forward under the context of the international project *Knowledge Dynamics in the Field of Social Sciences: Abduction, Intuition and Invention* (CFCUL/Universidade de Sevilha). Let us express our warmest thanks to all the participants to that event for fruitful and thought-provoking discussions.

Let us too express our thanks to all the reviewers of the chapters, whose thorough work definitely improved the final text.

Contents

Part I
The Dynamics of Knowledge I: Proof-Theoretical Approaches and the Interactive Viewpoint

Chapter 1
Perennial Intuitionism

Johan G. Granström

Abstract The basic tenets of intuitionism are the rejection of the law of excluded middle and the view that a judgement is correct if it is knowable, indicating a reversal in priority between the objective and the subjective. Intuitionism revived the age-old problem of universals, and the controversy between nominalism, conceptualism, and realism, now represented by formalism (nominalism), intuitionism (conceptualism), and set-theoretical Platonism (realism). In the old controversy, moderate realism, i.e., the Aristotelic-Thomistic school, came out on top, with its simultaneous rejection of conceptualism and exaggerated realism, on the grounds that the former leads to subjectivism, and the latter is epistemologically untenable. This paper takes a similar stance in the modern foundational debate: set-theoretical Platonism, is rejected on epistemological grounds, and pure conceptualism is rejected on the grounds that if fails to account for the objective nature of mathematics.

Keywords Intuitionism • Intuitionistic logic • Foundations of mathematics • Philosophia perennis • Formalism • Platonism • Nominalism • Conceptualism • Realism • Law of excluded middle • Bivalence • Epistemology

1.1 Prolegomena

The word logic, or, rather, its Early English spelling *logike*, is a direct transliteration of the word λογική, first used in its present sense by Zeno the Stoic.[1] The word λογική is in turn derived from the word λόγος with a wide range of meanings from the concrete, word or speech, to the abstract, discourse or reason.[2] According to ancient philosophers, concepts are derived from things and words are expressions

[1] Cf., Diogenes, *Lives of Eminent Philosophers*, Ch. 7, in particular n. 32, sqq.; and Cicero, *De Fato*, n. 1.

[2] From *oratio* to *ratio*, to use two common Latin translations of the word λόγος.

J.G. Granström (✉)
Kreuzstrasse 25B, CH-8802 Kilchberg, Schweiz
e-mail: georg.granstrom@acm.org

J. Redmond et al. (eds.), *Epistemology, Knowledge and the Impact of Interaction*,
Logic, Epistemology, and the Unity of Science 38, DOI 10.1007/978-3-319-26506-3_1

of thoughts.[3] The classical view on this threefold correspondence is that things have priority over thoughts, and thoughts over words, as eloquently expressed by Cajetan in the beginning of his commentary of Aristotle's *Categories*[4]:

> And even if we have to maintain this interpretation of the intention of this book, we must not forget what Avicenna so aptly says at the beginning of his Logic, namely, that to treat of words does not pertain to logical discussions on purpose, but it is only a sort of necessity that forces this on us, because the things so conceived we cannot express, teach, unite, and arrange, but by the help of words. For if we were able to carry out all these things without the use of external words, satisfied by the use of internal speech alone, or if by other signs would these things be achieved, it would be pointless to treat of words. So if one were to ask whether it is words or things which are principally treated of here, we have to say that it is things, though not absolutely, but insofar as they are conceived in an incomplex manner, and, by consequent necessity, insofar as signified by words.[5]

Cajetan mentions that it is a sort of necessity which forces the treatment of words upon us because, if a thought is to be communicated, there has to be words for it. This insight is a kind of contrapositive to Wittgenstein's famous dictum: "Whereof one cannot speak, thereof one must be silent."[6]

As a general rule, the more experienced we are in a particular field the less we pay attention to the signs and expressions of the field and even to their meanings: instead our attention is entirely focused on the things (Fig. 1.1).[7] As an example, consider the driver of an automobile approaching a stop sign. The experienced driver does not pay attention to the word stop, nor to the red colour or to the hexagonal shape; perhaps he does not even become conscious of the significance of the sign—he

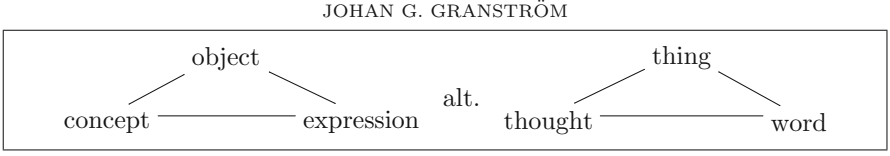


Fig. 1.1 The relation between object, concept, and expression: and the old-fashioned triple: thing, thought, and word

[3] Cf., Aristotle, *Perih.*, Ch. 1

[4] To aid the understanding of the first part of this quotation, it should be added that Cajetan's interpretation of Aristotle's point of view is that words are signs of concepts and that concepts are signs of things (cf., ibid., Ch. 1, 16a4).

[5] Cajetan, *In Praed.*, Ch. 1

[6] Wittgenstein, *Tractatus*, § 7: "Wovon man nicht sprechen kann, darüber muss man schweigen." The exact contrapositive of Cajetan's point is that, if there are no words for a thought, then it cannot be communicated. This is tantamount to Wittgenstein's dictum.

[7] Note that 'On Concept and object' is Geach's translation of the title of Frege's article 'Über Begriff und Gegenstand'. Cf., Maritain, *The Degrees of Knowledge*, Ch. 3, § 10, § 24; De Morgan, *Formal Logic*, Ch. 2; Husserl, *Log. Unt. II*, Pt. 1, Inv. 1, § 33.

simply stops, habitually, as it were. In a like manner, the scientist learns to see through the expressions of his field and, to a certain extent, even their meanings.[8] This is all well, except in philosophy and logic, where we *have to see* the words and their meanings to be able to investigate them.

One way of dividing logic is according to the three acts of the mind: simple apprehension, judgement, and reasoning (Table 1.1).[9] Simple apprehension, or perception, is an act of the mind in which the intellect comes to know something, as, for example, through sight. The detailed study of apprehension belongs to psychology, but, the existence of this act is of importance also to logic, since it provides the mind with raw material about which to think. A judgement is defined as an act in which the intellect recognises some form of agreement or discrepancy between concepts. Reasoning, treated of after the judgement, is an act of the mind by which, from known premisses, the mind comes to know a conclusion.

The main stream of logic has gradually turned from the principally material logic of the scholastic period to the prevailing formal logic,[10] through the influence of logicians such as Leibniz, Boole, and Frege, culminating in the formalistic crown jewel *Principia Mathematica*, by Whitehead and Russell, published in 1910.[11] But a complete method of logic must account both for the formal side of logic, i.e., how concepts are expressed, and for the material side of logic, i.e., how things are conceptualised. Instead of formal and material, one could use the modern counterparts syntactic and semantic: thus, we speak of a formal-material or syntactic-semantic method of logic.[12]

An expression which consists of a single meaningful word is called a categorem, and a word which is meaningful only in combination with other expressions is called a syncategorem. For example, in the expression

two plus three times five,

[8] Another example, due to Descartes, is that it may happen that we remember something that somebody told us, without remembering in which language it was spoken ('The World or Treatise on Light', Ch. 1, n. 4).

[9] Author's translation of a table from the Editors' preface of Aquinas, 'In Perih.', p. ix. Simple apprehension is the scholastic term used, e.g., by Gredt, *Elem. Phil.* n. 6. Perception is a modern equivalent used, e.g., by Locke, *An Essay Concerning Humane Understanding*, Bk. 2, Ch. 9. Cf., Arnauld, *The art of thinking*, p. 29, Kant, *Kritik der reinen Vernunft*, and Bolzano, *Wissenschaftslehre*.

[10] Cf., Bocheński, *Ancient Formal Logic*.

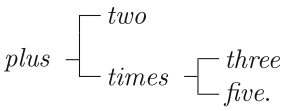
[11] Since the definition of logic is a controversial matter, we have avoided it completely. Cf., Husserl, *Log. Unt. I*, § 3; Mill, *A System of Logic*, § 1; and Gredt, *Elem. Phil.* n. 4.

[12] The syntactic-semantic method of logic is associated with Martin-Löf. It should be noted that the words formal and syntactic also have modern senses, originating with Hilbert and Carnap respectively, according to which only that is formal or syntactic which treats of words without regard to meaning or content.

Table 1.1 A classical division of logic, as it appears in the works of Aristotle, according to the acts of the mind

Logic	(A) Acts of the mind by which something is understood, two in number:	(1) *Simple apprehension*, which is treated of by the doctrine handed down from Aristotle in the *Categories*;	
		(2) *Judgement*, in which is truth or falsity, which is treated of by the doctrine handed down from Aristotle in the book *Perihermeneias*.	
	(B) Acts of the mind by which reasoning proceeds from one to another, as regulated by *logic*. This act has three moods of procedure by which it deduces the conclusion, either necessary, probable, or false.	(1) *Analytic* or *judgemental logic*, which proceeds by resolution, treats of the minds moods of procedure which *induce necessity*, and can be considered in two ways:	(a) either from the point of view of the *form* of the syllogism, as is done in the book *Analytica Priora*;
			(b) or from the point of view of the *matter* of the syllogism, as is done in the book *Analytica Posteriora*.
		(2) *Inventive Logic* treats of the minds moods of procedure which *induce probability*, and is divided into three, according to what it generates: faith, suspicion, or appreciation:	(a) in *faith* and *opinion*, the mind is totally inclined towards one of two contradictories, but allows with dread for the other: and to this pertains the *Topics*;
			(b) in *suspicion*, the mind is not totally inclined towards either contradictory: and to this pertains the *Rhetoric*;
			(c) in *appreciation*, the soul is inclined towards one of the two contradictories because of its beautiful representation: and to this pertains the *Poetics*.
		(3) The part of logic which is called *Sophistry* treats of the minds moods of procedure which *induce error*, and Aristotle treats of this in the book *Elenchorum*.	

the words *plus* and *times* are syncategorems and the words *two*, *three*, and *five* are categorems. This structure becomes apparent if the expression is displayed in tree-form:



The categorems are the leaves of the tree and the syncategorems are the internal nodes.

The only thing demanded of categorems and syncategorems is that they be recognizable as instances of some abstract *form*. The categorems and syncategorems are words, printed on paper or spoken out loud, but their meanings are not in the concrete words but in the abstract forms to which we recognise that the words belong.[13] In the example above, *plus* is the form, with *two* and *three times five* as parts; continuing the analysis, *two* is a form without parts and *three times five* has *times* as form and *three* and *five* as parts.

Now we have the terminology in place to spell out the *principle of compositionality*: the meaning of a complex expression is determined by the meanings of its parts, together with a meaning contribution from the form.[14] Expressions can be either simple or complex and, according to the principle of compositionality, a complex expression has a complex meaning. Note that the principle of compositionality says nothing about the converse. Thus, a categorem may well have complex meaning: this is the case, it seems, when we make abbreviatory definitions.

1.2 Truth and Knowledge

Intuitionistic type theory is not in conflict with common sense realism even though the former, *prima facie*, seems to be a conceptualist framework. When the ancients spoke about objects and propositions, they had in mind *men*, *horses*, and *this man is sitting on the horse*. When modern philosophy speaks about objects and propositions, it has in mind *numbers*, *primes*, and *this number is prime*. In ancient and medieval philosophy the focus is on real things and the treatment of mathematical entities is often a kind of appendix, whereas in modern philosophy it is typically the other way around. Since intuitionistic type theory is supposed to be able to account for propositions concerning both the real and the ideal, this tension has to be relieved.[15]

The meaning of an expression is the concept expressed by it and the referent of a concept, or of its expression, is the object signified. It is primarily the concepts which refer to their objects; the expressions refer only in a secondary sense: "an expression only gains an objective reference *because* it means something, it can rightly be said to signify or name the object *through* its meaning."[16] Sometimes the word denotation is used as synonymous with reference and an expression is said to

[13] In 'The theory of algorithms', nn. 5–7, p. 2, Markov makes the same distinction between what he calls *elementary signs* and the corresponding forms, which are called *abstract elementary signs*.

[14] This principle is commonly attributed to Frege, even though it was not explicitly formulated by him.

[15] Cf., Cocchiarella, 'Conceptual Realism as a Formal Ontology'.

[16] Husserl, *Log. Unt. II*, Pt. 1, Inv. 1, § 13 (Author's translation). Cf., the parallel place in Aquinas, 'Summa Theol.', Pt. 1, q. 13, a. 1: "voces referuntur ad res significandas, mediante conceptione intellectus": words refer to things signified, through the intellect's concept (Author's translation).

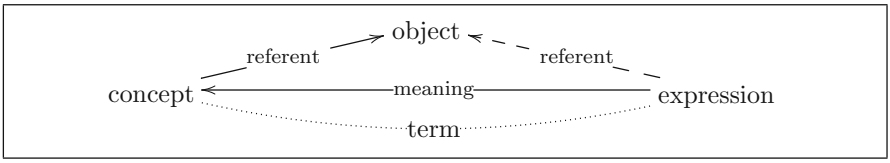


Fig. 1.2 Meaning, referent, and term added to the triangle picturing the threefold correspondence between object, concept, and expression

stand for, signify, or name its referent. In the case of universal concepts, which refer to many objects, the objects are said to *fall under* the concept.

In addition to expression, meaning, and concept, logic also uses the word term.[17] Is the word term to be identified with expression, concept, or object? The classical definition of a term is that into which a predication can be analysed, namely, the predicate and the subject.[18] In the classical literature, the word term is used ambiguously between the expression and its meaning and, when a clarification is called for, the scholastic authors write *terminus scriptus* for the expression and *terminus mentalis* for the concept. In my opinion, the best way to understand the word term is as *an expression taken together with its meaning* (Fig. 1.2). That is, it is neither the expression nor the concept, but both expression and concept taken together with the relation between them, i.e., the *meaningful expression*. One consequence of this is that, for two terms to be equal, they have to have the same unambiguous expression. For example, even if the words *freedom* and *liberty* have the same meaning, they are considered distinct as terms. Similarly, the word *light* as used in *a light feather* and *a light blue colour* stands for different terms.[19]

Terms sometimes refer to their objects *through another concept*. Compare for example *Paris* and *the capital of France*. The *meanings* of these two expressions are certainly very different. Let us agree to call a concept *immediate* if it signifies its object without any intermediate concept, such as *Paris*, and *mediate* if it signifies its object through some intermediary, as is the case with *the capital of France*.[20] In intuitionistic type theory, a similar distinction is made between *canonical* and *noncanonical* terms or expressions.[21] In this setting, canonical corresponds to immediate and noncanonical to mediate: the names canonical and noncanonical

[17]The Greek word ὅρος became *terminus* in Latin.

[18]Aristotle, *An. Pr.*, Bk. 1, Ch. 1, 24b17. Cf., Boëthius, 'De syllogismo categorico'.

[19]This use of the word *term* can be motivated as follows: the ancients speak about the three *terms* of a syllogism; *equivocation* is the fallacy of using an equivocal middle term in a syllogism, as in the argument "no light is dark; all feathers are light; therefore, no feathers are dark" (by Celarent); "an utterance is not called equivocal because it signifies many external things but because in signifying those many external things, there correspond to it different concepts in the soul." (Buridan, *Summulae de Dialectica*, Treatise 3, Ch. 1, § 2).

[20]Cf., Gredt, *Elem. Phil.* n. 16a.

[21]Cf., Martin-Löf, *Intuitionistic Type Theory*, p. 7.

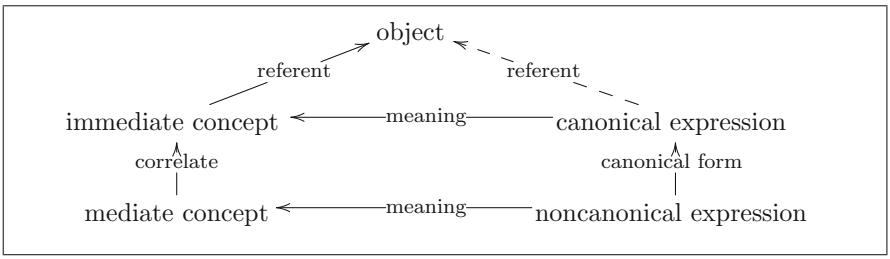


Fig. 1.3 Mediate and immediate concepts compared with canonical and noncanonical expressions

apply to terms and expressions while mediate and immediate apply to concepts. The canonical term which corresponds to a noncanonical term is called its *value*. There seems to be no established terminology for the relation between an immediate concept and the corresponding mediate concept, so we will call the immediate counterpart of a mediate concept its *correlate*. The whole picture is given in Fig. 1.3. Of course, the referent of a mediate concept is the same as the referent of its correlate, and the referent of a noncanonical expression is the same as the referent of its canonical form, which is the same as the referent of its meaning, i.e., of its concept.[22]

A correct understanding of the notion of *concept* is necessary for the correct understanding of the notion of judgement. Here, the doctrine of the concept as a *formal sign* is useful.

> A formal sign is a sign whose whole essence is to signify. It is not an object which, having, first, its proper value for us as an object, is found, besides, to signify another object. Rather it is anything that makes known, before being itself a known object. More exactly, let us say it is something that, before being known as object by a reflexive act, is known only by the very knowledge that brings the mind to the object through its mediation.[23]

> We signify our concepts to others by spoken words. And that is so because in order to make known to others the very objects we know, we communicate to them the same means, the same formal sign, that we ourselves use to know these objects.[24]

It is with knowledge of concepts as with knowledge of grammar: they can be known on two levels. The concept can be known "by the very knowledge that brings the mind to the object through its mediation" just as grammar can be known as proficiency in the art of grammar. On the second level, the concept can be known "by a reflexive act", in the same way as grammar can be known through explicit

[22]To use the jargon of mathematical category theory, the diagram presented in Fig. 1.3 is *commutative*, i.e., following any chain of arrows from one point to another gives the same result.

[23]Maritain, *The Degrees of Knowledge*, Ch. 3, § 24.

[24]Ibid., App. 1, § 4, p. 419. That is, communication does not only consist in an exchange of words, but also of their meanings.

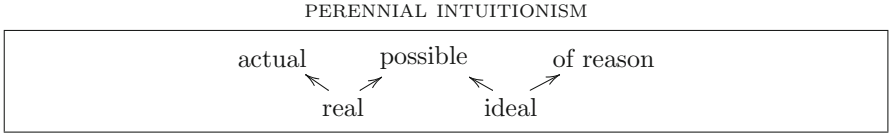


Fig. 1.4 Division of modes of being: actual, possible, and of reason, into real and ideal

knowledge of its laws, i.e., as a science. In the former case we speak about a *direct concept* and in the latter case we speak about a *reflex concept*.[25]

Concepts normally refer to things: thus, the first division of concepts is according to the things they refer to. Some things actually exist, such as animals and trees.[26] Other things have only possible existence, such as *a building five feet higher than the highest building in the world*. Moreover, things which have been actual, such as *a mammoth* or *Socrates*, are also called possible. Such things, possible or actual, are collectively called real beings—either because they are actual, because they can become actual, or because they have been actual.[27]

Yet another kind of being is that which is called a being of reason. Beings of reason cannot correspond to any thing, i.e., they cannot have any object *a parte rei*,[28] but exist only in the mind: "we say that these exist in the mind because the mind busies itself with them as kinds of being while it affirms or denies something about them".[29] Merely possible beings and beings of reason are collectively called ideal beings. Thus, an ideal being *does not* exist, whereas a being of reason *cannot* exist. The complete picture is given in Fig. 1.4.[30] Note that a possible being is called both real and ideal.[31]

For example, blindness is a being of reason. To be blind means not to have sight. The concept blindness is formed from the concept sight by adding negation. Similarly with death, deafness, and other privations. Another kind of beings of reason are those which are a result of a formal abstraction, such as the line or circle of geometry or the numbers of arithmetic, which are totally devoid of sensible matter and thus cannot exist in physical reality. Other examples of formal abstraction

[25]Poinsot, *Material Logic*, p. 421. Cf., Husserl, *Log. Unt. II*, Pt. 1, Inv. 1, § 34; and Gredt, *Elem. Phil.* n. 16c.

[26]Cf., Aristotle, *Metaph.*, Bk. 7, Ch. 1, for the various senses of the word *being*.

[27]Strictly speaking, things which are actual are also called possible (*ab actu ad posse valet illatio*) so *real being* and *possible being* amount to the same; but, when real being is divided into actual and possible, possible has to be taken to exclude actual.

[28]*A parte rei*: on the side of things.

[29]Aquinas, *In Metaph.*, Bk. 4, Les. 1, n. 12: "quam dicimus in ratione esse, quia ratio de eis negociatur quasi de quibusdam entibus, dum de eis affirmat vel negat aliquid" (trans. Rowan).

[30]After Maritain, *The Degrees of Knowledge*, Ch. 2, fn. 43.

[31]Cf., Husserl, *Log. Unt. II*, Pt. 1, Inv. 1, § 32: "Ideality in the ordinary, normative sense does not exclude reality" (trans. Findlay).

are the formation of the concept *redness* from *red*, *humanity* from *man*, etc. Yet another kind of being of reason are those of grammar and logic, such as *subject*, *predicate*, *proposition*, *set*, and *element*. Beings of reason are purely meaningful, or intelligible, entities, for which *the definition is everything*. How does the doctor confirm blindness in a patient? He checks for sight and when he does not find it he concludes blindness.

Of course, different beings of reason can be more or less distant from what is real. For example, a particular blindness is more real, more tangible, than, say, a particular prime number. In this sense, beings of reason admit of degrees in their distance from the real.[32] The Peripatetics maintain that the most basic concepts come from the direct apprehension of real being to the point that the nature, or species, of the thing is identified with the concept. This is the origin of the scholastic term *species expressa* for the concept. This nature, which, in a sense, is identified with the concept, is explained as follows by St. Thomas:

> Therefore, if it is asked whether this nature considered in this way can be said to be one or many, neither alternative should be accepted, because both are outside of the understanding of humanity, and either can pertain to it. For if plurality were included in its understanding, then it could never be one, although it is one insofar as it is in Socrates. Likewise, if unity were included in its notion and understanding, then Socrates and Plato would have one and the same nature, and it could not be multiplied in several things.[33]

The analogy of a work of art will make this view clearer. Consider, e.g., Homer's *Iliad*. If you buy a copy of the *Iliad* in a bookstore, you get a piece of matter, paper, along with it, and, necessarily so, since the work itself cannot be communicated but by the help of matter. Thus, the *Iliad* exists in your copy in the same way as the nature of a tree exists in the tree. Moreover, in one sense, it is the same work in different copies, and, in an analogous sense, it is the same nature in different trees (Fig. 1.5).

According to the ancients, mathematical concepts, such as number, line, triangle, etc., are a result of formal abstraction from sensible matter, according to the Peripatetic axiom *nihil est in intellectu quod non fuerit prius in sensu*.[34] In a formal abstraction we disengage from the real, and ideal mathematical entities are founded on the real in the sense that they are the result of a formal abstraction from it. However, these are also disengaged from the real in the sense that, e.g., the *mathematical definition of number* does not contain any reference to reality.

---

[32] Cf., Maritain, *The Degrees of Knowledge*, Ch. 2, § 33, p. 144

[33] Aquinas, 'De ente et essentia', Ch. 2: "Unde si quaeratur utrum ista natura sic considerata possit dici una vel plures, neutrum concedendum est, quia utrumque est extra intellectum humanitatis et utrumque potest sibi accidere. Si enim pluralitas esset de intellectu eius, nunquam posset esse una, cum tamen una sit secundum quod est in Socrate. Similiter si unitas esset de ratione eius, tunc esset una et eadem Socratis et Platonis nec posset in pluribus plurificari." (Trans. Klima).

[34] Aquinas, 'De Veritate', q. 2, a. 3, arg. 19. Author's translation: nothing is in the intellect that was not previously in the senses. Cf., Coffey, *The Science of Logic*, p. 7.

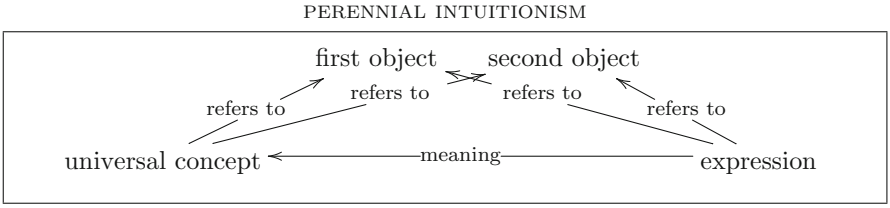


Fig. 1.5 The threefold correspondence for universal concepts, which refer to many objects. In this figure there are two objects, but there can be arbitrarily many

This ideal quality of the mathematical concept of number should not lead us to believe that the connection to reality is of no importance. Take for example Lagrange's four square theorem, that every natural number n can be written as a sum of four squares

$$n = a^2 + b^2 + c^2 + d^2.$$

Having demonstrated this theorem, we want to be sure that all gravel in the nearest gravel-pit can be divided into four piles, each of which can be laid in a square. It is so because number is a being of reason founded on real being. Thus, while intuitionistic type theory strictly speaking deals only with beings of reason, these beings of reason have to be founded on real beings: otherwise the whole project is reduced to inanity or mere navel-contemplation.

As said above, beings of reason do not have any object *a parte rei*. If we speak of an object for them, it is a purely formal or mathematical object (Fig. 1.6). In this precise sense, intuitionistic type theory can be labelled a conceptualist framework. But, as is clear from the above, there is no conflict between conceptualism for beings of reason and common sense realism. Thus, with respect to the age-old controversy between realists and conceptualists,[35] the present approach intuitionistic type theory should be acceptable to both parties, as realists agree that beings of reason have no object *a parte rei*.

The rejection of Platonic objects with extra mental existence does not make beings of reason into something subjective. Two senses of the word *objective* can be distinguished: the first and primary sense of the word is *on the object side* of the triangle; the second and derived sense is *the opposite of subjective*; it is derived because real things are not subjective. The second sense of the word objective is better described by the word *transsubjective*,[36] and mathematics is objective in this second sense, but not in the first sense, since its objects are formal, i.e., it mathematics does not have real being at the object vertex of the triangle.

[35]Cf., e.g., Gredt, *Elem. Phil.* n. 114.

[36]Cf., Husserl, *The Crisis of European Sciences and Transcendental Phenomenology*, Pt. 2.

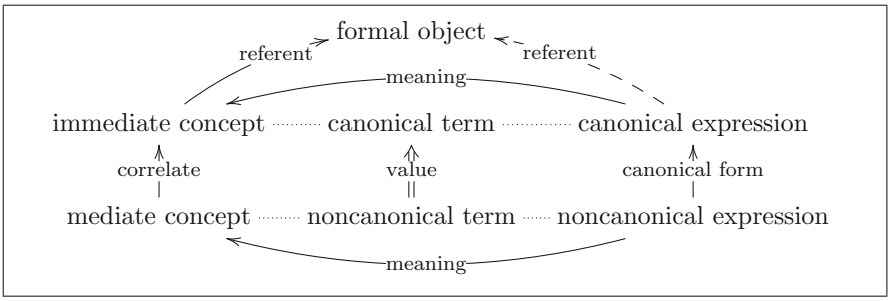


Fig. 1.6 For beings of reason we get a threefold correspondence between expression, concept, and formal object

In fact, all well-defined beings of reason are objective in the second sense because they are firmly founded in intelligible relations between concepts, or, to use Biancani's term, in intelligible matter.[37] The whole of mathematics and intuitionistic type theory serve as examples of this objectivity. Husserl makes it clear that we can speak of meanings in themselves,[38] and the scholastic counterpart of these meanings in themselves is the *conceptus objectivus*, i.e., the concept taken in its objective, as opposed to mental, aspect.[39] When philosophers say that mathematics is founded in intelligible matter, they mean precisely that the formal object, the meaning in itself, or the objective concept, is objective in the second of the above two senses, i.e., transsubjective.[40]

This brings us to the important question of mathematical existence. That a certain expression is meaningful does not guarantee that its formal object *exists*. Husserl makes a distinction between *nonsense* and *absurdity*.[41] For example, we call *a largest prime number*, and *a square circle* absurd, but these expressions are still meaningful, i.e., they have a sense. If they did not have a sense we could not say that they do not exist.[42] Of course, the above manner of speaking about intelligible matter does not settle what makes an ideal object existent, as opposed to absurd.[43] There are three main answers to this question: Platonism, formalism, and intuitionism.

[37]Blancanus, 'A Treatise on the Nature of Mathematics along with a Chronology of Outstanding Mathematicians', pp. 179–180.

[38]Husserl, *Log. Unt. II*, Pt. 1, Inv. 1, § 35.

[39]Cf., Gredt, *Elem. Phil.* n. 7.

[40]Cf., Maritain, *The Degrees of Knowledge*, Ch. 4, § 6, pp. 152–154.

[41]Husserl, *Log. Unt. II*, Pt. 2, Inv. 6, § 12. Cf., ibid., Pt. 1, Inv. 1, § 15.

[42]Maritain calls absurd beings of reason the "thieves and forgers" among beings of reason (*The Degrees of Knowledge*, Ch. 2, § 33, p. 143).

[43]Cf., Bernays, 'Mathematische Existenz und Widerspruchsfreiheit'.

Platonism. The question of existence does not pose much of a problem for Platonism as, according to this doctrine, the mathematical entities are as real as horses and elephants.[44] On the other hand, the distinction between *nonsense* and *absurdity* becomes problematic since, in its extreme form, Platonism is bound to claim that everything which is absurd is also nonsense: if sense entails existence, then lack of existence, i.e., absurdity, entails lack of sense. For example, it would be necessary to reject a geometrical figure that is both square and round as nonsense, since it does not exist.

Formalism. Formalism is associated with the idea that, if an object can be spoken about consistently, then it exists.[45] This view is motivated by certain mathematical insights of historical importance, e.g., that one can consistently add negative and irrational numbers to the language of arithmetic, because anything that can be demonstrated using them can also be demonstrated without using them. An objection against this view is that, for real being, consistency does not entail existence, so why should it do so for beings of reason? For example, one can consistently assume that there is intelligent life on another planet, since this assumption will never be refuted; but this does not entail that such life exists in the usual sense of the word.[46]

Intuitionism. Before treating of intuitionism, it is instructive to consider the point of view of finitism. The basic tenets of finitism are that all of mathematics should ultimately be founded on the natural numbers and that the natural numbers themselves are founded on the numerals.[47] Intuitionism is a refinement of finitism that adds an important ingredient, namely, the notion of *mental construction*.[48] Mathematical objects are not conceived as pre-existing and singled out by descriptive definitions: instead they are constructed mentally, thereby avoiding the existence problem.

Several prominent mathematicians and philosophers have taken part in the debate between the three main schools.

> Frege: "This is the predicament of formal arithmetic: it cannot help but make use of sentences supposed to express thoughts, but nobody can determine exactly what these thoughts are."[49]

[44]Cf., Maddy, 'Mathematical existence'.

[45]Hilbert, 'On the infinite', p. 370. Cf., what von Neumann reportedly said to a colleague who didn't understand the method of characteristics: "Young man, in mathematics you don't understand things. You just get used to them."

[46]Cf., Becker, 'Mathematische Existenz'.

[47]This view was expressed by Kronecker in the famous sentence "Die ganze Zahl schuf der liebe Gott, alles übrige ist Menschenwerk" (Cajori, *A History of Mathematics*, p. 362).

[48]This notion was introduced by Brouwer. Kant is the likely source of his terminology: "Philosophical knowledge is the knowledge gained by reason from concepts; mathematical knowledge is the knowledge gained by reason from the *construction* of concepts." *Kritik der reinen Vernunft*, Pt. 2.1.1, p. 469 (B 741) (trans. N. K. Smith).

[49]Frege, *Grundgesetze der Arithmetik II*, § 105 (trans. Black).

Weyl: "If Hilbert's view prevails over intuitionism, as appears to be the case, *then I see in this a decisive defeat of the philosophical attitude of pure phenomenology*".[50]

Bourbaki: "The intuitionist school, whose memory will undoubtedly survive only as a historical curiosity, has at least rendered the service of having obliged its opponents, that is to say the vast majority of mathematicians, to clarify their own positions and to become more consciously aware of the reasons (whether logical or sentimental) for their confidence in mathematics."[51]

Simpson: "We have mentioned three competing 20th century doctrines: formalism, constructivism, set-theoretical Platonism. None of these doctrines are philosophically satisfactory, and they do not provide much guidance for mathematically oriented scientists and other users of mathematics. As a result, late 20th century mathematicians have developed a split view, a kind of Kantian schizophrenia, which is usually described as "Platonism on weekdays, formalism on weekends". In other words, they accept the existence of infinite sets as a working hypothesis in their mathematical research, but when it comes to philosophical speculation, they retreat to a formalist stance. Thus they have given up hope of an integrated view which accounts for both mathematical knowledge and the applicability of mathematics to physical reality. In this respect, the philosophy of mathematics is in a sorry state."[52]

Skolem: "I believed that it was so clear that axiomatization in terms of sets was not a satisfactory ultimate foundation of mathematics that mathematicians would, for the most part, not be very much concerned with it. But in recent times I have seen to my surprise that so many mathematicians think that these axioms of set theory provide the ideal foundation for mathematics; therefore it seemed to me that the time had come to publish a critique."[53]

Bishop: "The fact that space has been arithmetized loses much of its significance if space, number, and everything else are fitted into a matrix of idealism where even the positive integers have an ambiguous computational existence. Mathematics becomes the game of sets, which is a fine game as far as it goes, with rules that are admirably precise. The game becomes its own justification, and the fact that it represents a highly idealized version of mathematical existence is universally ignored."[54]

1.3 Judgement and Demonstration

From grammar, we learn that a sentence is the verbal, oral or written, expression of a complete thought.[55] In logic, the word assertion is used for a sentence susceptible of logical analysis, and the word judgement for its mental counterpart. We now come to a crucial point, namely the notions of *correctness* and *evidence* for judgements

[50] Weyl, 'Comments on Hilbert's second lecture on the foundations of mathematics', p. 484.

[51] Bourbaki, *Elements of Mathematics*, p. 336.

[52] Simpson, 'Logic and mathematics', § 3.2.

[53] Skolem, 'Some remarks on axiomatized set theory', pp. 300–301.

[54] Bishop et al., *Constructive Analysis*, Ch. 1, p. 7.

[55] Sentence is that which was λόγος in Greek and became *oratio* in Latin.

and assertions.[56] How to reconcile the Aristotelic-Thomistic theory of truth, or correctness, of judgements with the intuitionistic view?

In the Methaphysics, we find the important remark that "falsity and truth are not in *things* but in *thought*".[57] A judgement is first and foremost an *act* and, as act, it has an agent: the content of an evident judgement is always evident *to* somebody; in which case it is nothing but a piece of his knowledge.[58] The content of a judgement can become evident through a variety of means, including apprehension, definition, and demonstration. Now, the controversial intuitionistic defintion of correctness of the content of a judgement is that the content is *correct* if it *can be made evident*. The virtue of this definition is that it includes all means through which a judgement can be made evident. Since evident and known are interchangeable, this definition can be paraphrased by saying that the content of a judgement is correct if it is knowable. As the content of the judgement is something objective, in the second of the above two senses, this definition indicates a reversal of priority between the objective and the subjective: the objective correctness is defined in terms of the subjective evidence.

The Aristotelic-Thomistic notion of truth can be summed up in two formulae: "to say that what is not, or what is not is, is false; but to say that what is, or what is not is not, is true"[59]; and the proverbial "truth is the adequation of thing and intellect".[60] These formulae both mention *things*, and, consequently, they deal with truth or correctness for judgements about reality. To reconcile them with the intuitionistic definition of correctness, it is sufficient to make explicit a rule that we take for granted, viz., that a judgement involving concepts derived from reality must have evidence drawn from reality. The conclusion is that evidence is conceptually prior to correctness, whereas reality is ontologically prior to any judgement about reality being evident.[61]

Reasoning is an act of the mind by which a certain judgement, the conclusion, is made evident: that is, the final act in a piece of reasoning is the act of judging its conclusion. The verbal expression of a piece of reasoning is called an argument, when dealing with reasoning in general, or a demonstration, when dealing with exact sciences.

[56] We prefer the word correct to the word true to avoid confusion with true propositions, discussed later.

[57] Aristotle, *Metaph.*, Bk. 6, Ch. 4, § 2. Cf., Moore, 'The nature of judgement', p. 179.

[58] Cf., Martin-Löf, 'On the meanings of the logical constants and the justifications of the logical laws', p. 24 and ibid., p. 19.

[59] Aristotle, *Metaph.*, Bk. 4, Ch. 7, § 1.

[60] Author's translation of "veritas est adaequatio rei et intellectus", Aquinas, 'Summa Theol.', Pt. 1, q. 16, a. 2.

[61] One concept is conceptually prior to another if the definition of the latter involves the former, and one thing is ontologically prior to another if the latter cannot be conceived as existing without the former existing also.

A demonstration is analysed into inferences.[62] The mental counterpart of an inference brings the mind from certain judgements already made, the premisses, to a new judgement, the conclusion, which becomes known. We will write inferences on the form

$$\frac{P_1 \quad \cdots \quad P_n}{C,}$$

where P_1 up to P_n are the premisses and C is the conclusion.

More geometrico, a demonstration must start from premisses which are immediately known, without any need for further demonstration. Such an assertion is called an axiom, ἀξίωμα in Greek. In addition, certain inference steps are *immediate*, i.e., they do not admit further analysis. Instead of immediate, which is something negative, i.e., the absence of a means, one could say self-evident.[63] Thus, an assertion or inference is self-evident if it is "known by reason of the terms themselves, or by the explanation of the terms".[64] Instead of self-evident, an axiom or immediate rule of inference can be said to be evident *ex vi terminorum*, i.e., by force of the terms, or, which amounts to the same, *per se nota*, i.e., evident through itself. For immediate inferences, this means that, when the premisses are known, nothing more is called for to come to know the conclusion.[65]

There may be some discourse which leads to the acceptance of a self-evident assertion or inference, viz., the explanation of the terms.[66] This discourse is of course not demonstrative in the above sense of the word, but it may be termed apodictic in the derived sense of being necessary and absolute.[67] On the other hand, not every assertion accepted without discourse is self-evident. For example, assertions involving faith in a credible witness are accepted without discourse, but still not self-evident.[68]

[62]Martin-Löf, 'A Path from Logic to Metaphysics'; Sundholm, 'Inference versus Consequence'. Cf., also Aristotle, *An. Pr.*, Bk. 1, Ch. 1; *An. Post.*, Bk. 1, Ch. 10; *Top.*, Bk. 1, Ch. 1.

[63]Cf., Aristotle, *An. Post.*, Bk. 1, Ch. 2; Aquinas, 'In An. Post.', Bk. 1, Lect. 5; Poinsot, *Material Logic*, p. 461.

[64]Ibid., p. 462.

[65]This last explanation of what constitutes an immediate inference is due to Sundholm, 'Inference versus Consequence', p. 35. As an aside, in contrast to Whitehead and Russell, we do not think that an axiom can be accepted on purely practical grounds (cf., *Principia Mathematica*, Intro., Ch. 2, § 7, p. 62). The argument that "things have been taught to be self-evident and have yet turned out to be false" (ibid.) has little force, since, clearly, they were not self-evident after all: *errare humanum est*.

[66]Cf., Aquinas, 'Summa Theol.', Pt. 1, q. 2, a. 1.

[67]Cf., Aristotle, *An. Pr.*, Ch. 1.

[68]Cf., Poinsot, *Material Logic*, p. 462.

1.4 The Proposition

In intuitionistic type theory, a distinction is made between an assertion and a proposition.[69] Although this distinction, *prima facie*, seems to be subtle and of little importance, it turns out to have far-reaching consequences. This distinction is most clearly seen by an example where a proposition occurs unasserted. Let A and B be propositions, e.g., *the moon is a cheese* and *the moon is edible* respectively. Then *if A then B* is a new proposition, and in asserting that it is true, neither A, nor B, is asserted to be true. This is clear by the example. Geach calls this observation the Frege point, since Frege stressed it and made it explicit in the *Begriffsschrift*.[70]

Before defining what it means for something to be a proposition and what it means for a proposition to be true, two forms of assertion must be introduced, namely, that A is a proposition, written

$$A : \text{prop},$$

and that a proposition A is true, written

$$A \text{ true} .$$

That A is true presupposes that A is a proposition, since before we can know that a proposition is true, we must know that it is a proposition. The logical connectives operate on propositions. That is, granted that A and B are propositions,

$$A \,\&\, B, \quad A \vee B, \quad A \supset B, \text{ and } \Lambda$$

are also propositions.[71] One of the first and most important tasks of intuitionistic type theory is to explain the two forms of assertion, as well as the meanings of

[69] Subsequently we will prefer the word assertion to the word judgement. This choice differs from that of Martin-Löf, 'On the meanings of the logical constants and the justifications of the logical laws', who chooses *judgement* as the primary word, but it agrees with that of Russell, e.g., 'The Theory of Implication', § 1.1.

[70] Geach, *Logic Matters*, p. 255. But, as pointed out by Klima, the Frege point was recognised long before Frege, for example, by Buridan, in *Summulae de Dialectica*, Treatise 5, Ch. 1, § 3, p. 308: "a syllogism has an additional feature in comparison to a conditional in that a syllogism posits the premises assertively, whereas a conditional does not assert them."

[71] These connectives are called conjunction, disjunction, implication, and *falsum* (or *absurdum*) respectively. The word connective applies strictly speaking only to the first three, since they connect A and B, but the meaning of the word is often extended to include *falsum* too (as well as negation and equivalence, see below). The symbol & is a ligature for the Latin word *et* meaning *and*; the symbol $\vee$ is just a stylised abbreviation of the Latin word *vel* meaning *or*; the symbol $\supset$ is due to Peano (*Arithmetices Principia Nova Methodo Exposita*, Log. Not., n. 2), in fact, $\subset$ is a stylised C abbreviating *is a consequence of*, so $B \subset A$ means that B is a consequence of A, or, equivalently, that A implies B; finally, the symbol Λ for *falsum* is due to Peano (ibid.), and it is a V for *verum* turned upside down.

these connectives, and the two quantifiers, in such a way as to make the laws of propositional and predicate logic evident.

In the traditional approach to logic, the first division of propositions is into affirmations and denials. Here the word proposition is used in its traditional sense, corresponding to what we call an assertion. This symmetric treatment of affirmation and denial goes back to Aristotle and is founded on the law of excluded middle.[72] The modern version of this symmetry is the interpretation of a proposition, now in the modern sense, as a truth value, i.e., as referring to *the true* or *the false*.[73] There are several problems with this symmetry between affirmation and denial.[74]

(a) The law of excluded middle is ontological, not logical. Bringing it into logic can be seen as an instance of the fallacy μετάβασις εἰς ἄλλο γένος.[75] I maintain that it is not a law of thought, i.e., a law of logic, but a principle of being.
(b) Although the law of excluded middle has a kind of intuitive validity for real being, it is not evident for beings of reason.[76] Should not the laws of logic hold for pure mathematics?
(c) Many predicates in natural language are vague and allow for borderline cases.[77] Such predicates do not fare well in classical logic but are treated of without problems in intuitionistic logic, where the law of excluded middle is not accepted as a law of thought.
(d) The laws for forming propositions by quantification over infinite domains are difficult to justify under the classical interpretation of a proposition as a truth value.[78]

So, what does intuitionism suggest instead of the definition of a proposition as a truth value? Put differently, what does the form of assertion A : prop mean?

Definition. A proposition is defined by laying down what counts as a cause of the proposition.

With this definition in place, it is natural to define truth of a proposition in the following way.[79]

[72] Aristotle, *Perih.*, Ch. 1 (cf., ibid., Ch. 4, 17a2).

[73] Boole, 'The Calculus of Logic'. Cf., Martin-Löf, 'On the meanings of the logical constants and the justifications of the logical laws', p. 14.

[74] Cf., Sundholm, 'Inference versus Consequence', p. 26.

[75] I.e., the jumping into a different domain or science. The phrase is derived from Aristotle, *An. Post.*, Bk. 1, Ch. 7, 75a38, which is concerned with the impossibility of proving facts in one science using the methods of another, e.g., to prove a geometrical fact by appeal to optics.

[76] Husserl, *Log. Unt. II*, Pt. 2, Inv. 6, § 30.

[77] Cf., Geach, 'The law of excluded middle', pp. 71–73.

[78] Martin-Löf, *Intuitionistic Type Theory*, p. 11, cf., Brouwer, 'The Unreliability of the Logical Principles'.

[79] These definitions are copies of Martin-Löf's definitions (*Intuitionistic Type Theory*, p. 11) with the word *proof* replaced by the word *cause*.

Definition. A proposition is true if it has a cause.

To understand the word *cause* in these definitions, consider the classical dictum *scire est rem per causas cognoscere*[80]: this notion of truth of a proposition has Leibniz's principle of sufficient reason, as it were, built in. The principle of sufficient reason is that "in virtue of which we hold that, no fact can be found true, nor can truth exist in any proposition, unless there be a sufficient reason, why it is so rather than otherwise, although these reasons most often cannot be known by us."[81]

Thus, when I say that I know that the proposition A is true, I mean that I am in possession of a cause of it. In this setting, the cause could also be called a *reason*,[82] i.e., the reason by which I know that A is true. The distinction between cause (*causa*) and reason (*ratio*) is a virtual distinction: a cause is taken as an *objective* ground of a proposition whereas a reason is taken as a *particular subject's* ground for holding the proposition true.[83]

These definitions, of proposition and truth, are of no value until it becomes clear that all classical laws of logic, except the law of excluded middle, can be justified from them by assigning suitable meanings to the logical connectives. The intuitionistic interpretation of the propositional connectives is given in Table 1.2.[84] Since a proposition is defined by laying down what counts as a logical cause of it (Table 1.2), the inference rules

$$\frac{A : \text{prop} \quad B : \text{prop}}{A \,\&\, B : \text{prop},} \qquad \frac{A : \text{prop} \quad B : \text{prop}}{A \vee B : \text{prop},}$$

[80]To know is to have cognizance of the thing through causes. This dictum is derived from Aristotle, *An. Post.*, Bk. 1, Ch. 2, 71b9, sqq. Cf., *Metaph.*, Bk. 2, Ch. 1, n. 5, sqq. Other formulations are the poetic "Felix, qui potuit rerum cognoscere causas" (Virgil, *Georgics*, Bk. 2, l. 490) and "Vere scire, esse per causas scire" (Bacon, *Novum Organum*, Bk. 2, Ch. 20). With respect to the division of causes (Aristotle, *Metaph.*, Bk. 5, Ch. 2; *Phys.*, Bk. 2, Ch. 3), the kind of cause we have in mind here could be called a *logical cause* (cf., *An. Post.*, Bk. 2, Ch. 11).

[81]Author's translation of Leibniz, 'Principia Philosophiæ', § 32: "vi cujus consideramus, nullum factum reperiri posse verum, aut veram existere aliquam enunciationem, nisi adsit ratio sufficiens, cur potius ita sit quam aliter, quamvis rationes istæ sæpissime nobis incognitæ esse queant."

[82]It is difficult to determine to what extent Leibniz identified *ratio* with *causa*; cf., Di Bella, 'Causa Sive Ratio'.

[83]In this setting, it also makes sense to call the cause or reason a *truth-maker*, since, in a sense, it is the cause that makes the proposition true. Cf., Sundholm, 'Existence, Proof, and Truth-Making: A Perspective on the Intuitionistic Conception of Truth'.

[84]Martin-Löf, *Intuitionistic Type Theory*, p. 12. This interpretation is called the BHK interpretation after its discoverers Brouwer (in many of his works), Heyting ('Sur la logique intuitionniste'), and, independently, Kolmogorov ('Zur Deutung der intuitionistischen Logik'). It should be mentioned that there is direct line of thought from Husserl to the BHK interpretation: Becker, one of Husserl's students, interpreted propositions as expectations ('Mathematische Existenz'), and influenced Heyting who interpreted propositions as problems (cf., Mancosu, *From Brouwer to Hilbert*, pp. 275–285). This leads to the identification of: (1) the cause of a proposition, (2) the fulfillment of an expectation, and (3) the solution of a problem.

Table 1.2 The intuitionistic interpretation of the propositional connectives, i.e., the BHK interpretation

A cause of	Consists of
$A \,\&\, B$	A cause of A and a cause of B;
$A \vee B$	A cause of A or a cause of B, together with information about which cause it is that is given;
$A \supset B$	A method which takes any cause of A into a cause of B;
Λ	(There is no cause of Λ)

and

$$\frac{A : \text{prop} \quad B : \text{prop}}{A \supset B : \text{prop}}$$

are self-evident, and so is the axiom

$$\Lambda : \text{prop}.$$

There are two connectives missing from this list, namely, negation and equivalence. These connectives can be defined in terms of the already introduced connectives by nominal definition. The negation of a proposition A is written $\sim A$ and defined by

$$\sim A \stackrel{\text{def}}{=} A \supset \Lambda : \text{prop}.^{85}$$

This definition of negation is commonly accepted in intuitionistic logic,[86] but other definitions have been proposed in other areas of logic. Equivalence between two propositions A and B is written $A \supset\subset B$ and defined by

$$A \supset\subset B \stackrel{\text{def}}{=} (A \supset B) \,\&\, (B \supset A) : \text{prop}.^{87}$$

This definition of equivalence seems to be universally accepted.

[85]The symbol $\sim$ for negation is due to Russell ('Mathematical Logic as Based on the Theory of Types', § 6).

[86]But, cf., Bishop et al., *Constructive Analysis*, pp. 10–11.

[87]The symbol $\supset\subset$ for equivalence is due to Heyting ('Die formalen Regeln der intuitionistischen Logik', § 2).

Having so defined the notion of assertion and explained the first two forms of assertion, namely, A : prop and A true, a distinction is to be made between a *complete* and an *incomplete* assertion.[88] The form of assertion

$$A \text{ true}$$

is incomplete in the sense that it suppresses the cause. We write

$$c : \text{cause}(A)$$

if c is a cause of A.[89] The meaning of the form of assertion A : prop is that it has to be laid down what counts as a cause of it. That is, a proposition A is defined by defining the form of assertion c : cause(A). Since that a proposition is true means that it has a cause, the inference rule

$$\frac{c : \text{cause}(A)}{A \text{ true}}$$

is self-evident and completely determines the meaning of the form of assertion A true.

The forms of assertion c : cause(A) and A : prop are both complete. Indeed, they are examples of the first form of complete assertion, the predication, where something (the predicate) is predicated of something (the subject). In intuitionistic type theory, the copula is often spelled colon which is read *is*.

Examples of predicates are 'prop' and 'cause(A)', for a proposition A. To get another example, define a number, in the sense of Peano,[90] to be either zero or the successor of a number. If we write 0 for zero and s(a) for the successor of a, we get the axioms

$$0 : \text{number}$$

[88] An incomplete assertion, e.g., A true, constitutes an *incomplete communication* (unvollständige Mitteilung) in that the speaker suppresses certain information (cf., Hilbert and Bernays, *Grundlagen der Mathematik*, p. 33; and Kleene, 'On the Interpretation of Intuitionistic Number Theory', § 1). Also, what we call an incomplete assertion was called a *judgement abstract* (Urteilsabstrakt) by Weyl ('Über die neue Grundlagenkrise der Mathematik', p. 54).

[89] This important step of bringing the causes into the language of logic, i.e., of naming them, was first taken by Martin-Löf, 'An intuitionistic theory of types', p. 77, under the guise of proof objects. Cf., Martin-Löf, 'Analytic and synthetic judgements in type theory', where the distinction between the complete assertion c : cause(A) and the incomplete assertion A true is related to the Kantian distinction between analytic and synthetic judgements.

[90] Peano, *Arithmetices Principia Nova Methodo Exposita*, § 1, with the difference that, as is now customary, the first number is zero instead of one. It is more natural to start the number series in the sense of Peano at zero since, if starting at one, there are two different formalizations of the unit, the starting point *one*, and the s for the successor.

and

$$\frac{a : \text{number}}{\text{s}(a) : \text{number.}} \tag{1.1}$$

This makes 'number' a third example of a predicate. Of course, propositions can involve numbers in the usual way. If $a < b$ is defined by stipulating that $a < \text{s}(a)$ for any number a, and that if $a < b$, then $a < \text{s}(b)$, then the inference rule

$$\frac{a : \text{number} \quad b : \text{number}}{a < b : \text{prop}} \tag{1.2}$$

becomes evident since $a < b$ is defined as a proposition.[91] Moreover the inference rules

$$\frac{a : \text{number}}{a < \text{s}(a)\ \text{true}} \tag{1.3}$$

and

$$\frac{a < b\ \text{true}}{a < \text{s}(b)\ \text{true}}$$

become evident in virtue of the definition.

The second form of complete assertion is the assertion of definitional equality. It turns out to be a bad idea to treat of equality in the general form

$$a = b,$$

because we first have to spell out what kind of objects a and b are, and, in this general form of equality, there is no guarantee that a and b have a common genus.[92] On the other hand, if we already know that $a : P$ and $b : P$ for some predicate P, then this form of assertion has good sense, and can be written

$$a = b : P$$

so as to explicitly show what kind of objects a and b are. That which stands on the right-hand side of the colon, i.e., the predicate P above, will be called a *logical*

[91] With mention of the causes, the definition of $a < b$ becomes: there is a cause of $a < \text{s}(a)$, and if there is a cause of $a < b$, then there is a cause of $a < \text{s}(b)$.

[92] Geach, 'Identity', p. 3. Cf., Quine's dictum: "no entity without identity" (*Theories and Things*, p. 102).

category.[93] In intuitionistic type theory, every logical category comes equipped with definitional equality. The relation of definitional equality between objects of any logical category should satisfy

(1) that the two terms of a definition are equal,
(2) that equals can be substituted for equals giving equal results,
(3) that any object is equal to itself, and
(4) that two objects which equal a third are equal to one another.[94]

An example of (3) is the assertion $0 = 0 : \text{number}$, and an example of (2) is the inference rule

$$\frac{a = b : \text{number}}{\text{s}(a) = \text{s}(b) : \text{number}.}$$

When defining things in the way we are used to in mathematics, we use definitional equality. For example, when addition between numbers is defined by the two equations

$$\begin{cases} a + 0 = a : \text{number}, \\ a + \text{s}(b) = \text{s}(a + b) : \text{number}, \end{cases}$$

the two sides of the equality sign are definitionally equal. To express this in inference rules, first note that the above definition of addition makes evident the inference rule

$$\frac{a : \text{number} \quad b : \text{number}}{a + b : \text{number},}$$

because $a + b$ can always be computed by the above equations. Moreover, the two inference rules

$$\frac{a : \text{number}}{a + 0 = a : \text{number}}$$

and

$$\frac{a : \text{number} \quad b : \text{number}}{a + \text{s}(b) = \text{s}(a + b) : \text{number}}$$

are evident from the definition of addition.

[93] See Klev, 'Categories and Logical Syntax' for a comprehensive analysis of the development of the notion of category from Aristotle to Kant, and beyond.

[94] Martin-Löf, 'About models for intuitionistic type theories and the notion of definitional equality', p. 93.

The definitional equality mentioned above is not the same as the usual mathematical equality. In mathematics, some equalities are definitional and some are not. When we prove some equality by mathematical induction, e.g., that addition is commutative,

$$a + b = b + a,$$

the equality is not definitional. This is because by proving it by induction we give the cause of the two terms being equal, i.e., this equality has to be expressed by a proposition. Consequently, a distinction has to be made between definitional equality and propositional equality: definitional equality is a complete form of assertion whereas propositional equality is a form of proposition. That two numbers a and b are propositionally equal will be written a eq b,[95] and this is a proposition, i.e.,

$$\frac{a : \text{number} \quad b : \text{number}}{a \text{ eq } b : \text{prop.}}$$

A cause of two numbers being propositionally equal is existent if they are definitionally equal, i.e.,

$$\frac{a = b : \text{number}}{a \text{ eq } b \text{ true .}}$$

That addition is commutative is now expressed by the incomplete assertion

$$(a + b) \text{ eq } (b + a) \text{ true},$$

i.e., the proposition $(a + b)$ eq $(b + a)$ is found to be true by finding a cause of it: in this case, the proof is by induction. Propositional equality can be negated, e.g., one of Peano's axioms for arithmetic is

$$\sim(0 \text{ eq } \mathrm{s}(0)) \text{ true .}$$

[95] We use the standard equality sign $=$ for definitional equality. This sign was introduced by Recorde, *The Whetstone of Witte*, in 1557: "And to avoide the tediouse repetition of these woordes : is equalle to : I will sette as I doe often in woorke use, a paire of paralleles, or Gemowe lines of one lengthe, thus: =, bicause noe 2 thynges, can be moare equalle." (there are no page numbers in this work, but the quoted passage stands under the heading "The rule of equation, commonly called Algebers Rule" which occurs about three quarters into the work). This use of the equality sign seems to me most natural since we use it when we make abbreviatory definitions in mathematics. Thus we had to use another sign for propositional equality. In the type-theoretic literature, there are several suggestions, including 'I' (Martin-Löf, *Intuitionistic Type Theory*, p. 59), and 'Id' vs. 'Eq' (with a slight difference in meaning, Nordström et al., *Programming in Martin-Löf's Type Theory*, Ch. 8). According to Cajori ('Mathematical Signs of Equality', p. 116), the most popular notation, both before Recorde and in competition with him, was to write equality in words, i.e., something like "æquales", "égale", "gleich", or the abbreviation "æq".

A proof of this axiom is given by Martin-Löf in *Intuitionistic Type Theory* (p. 91).[96] Since an assertion cannot be negated, this shows another difference between propositional and definitional equality.

A form of assertion typically has a number of *presuppositions* which are assertions which must be known in order for it to make sense.[97] In everyday language, presuppositions are most easily observed in sophistical questions, such as *Do you still beat your wife?*, which can neither be affirmed, nor denied, unless the presupposition is fulfilled. We have already seen that the forms of assertion A true and $c : \text{cause}(A)$ presuppose that A is a proposition; and that $a = b$: number presupposes that a and b are numbers. We will also use the word presupposition in a more general sense, according to which more specific forms of assertion can have more specific presuppositions. For example, that $A \ \& \ B$ is a proposition *presupposes*, in this more general sense, that A and B are propositions, because one cannot come to know that $A \ \& \ B$ is a proposition except by first knowing that A and B are propositions. Strictly speaking, inference rules where the conclusion is a presupposition of the premiss, in either sense, like

$$\frac{A \text{ true}}{A : \text{prop}} \quad \text{and} \quad \frac{A \ \& \ B : \text{prop}}{B : \text{prop},}$$

are *valid* but *useless*; the conclusion is already known before the premiss, so there is no point in inferring it.

An inference rule is called *well-formed*, if all presuppositions of the conclusion C can be inferred from the premisses P_1 up to P_n taken together with their presuppositions. When accepting a premiss, one also implicitly accepts its presuppositions, and if the presuppositions themselves have presuppositions, these are also accepted, etc. The relation of well-formedness imposes an order on the inference rules, because the validity of other inference rules may be needed to show that a particular inference rule is well-formed. For example, the conclusion of inference rule (1.3) presupposes that $a < \text{s}(a)$ is a proposition; this is demonstrated from the premiss a : number by

$$\frac{a : \text{number} \quad \dfrac{a : \text{number}}{\text{s}(a) : \text{number}}\ (1.1)}{a < \text{s}(a) : \text{prop};}\ (1.2)$$

thus, inference rules (1.1) and (1.2) have to come before inference rule (1.3).

With respect to the relation between demonstrability and correctness of an assertion, two questions can now be formulated.[98]

[96] Thus, properly speaking, this is not an axiom, but a theorem, of intuitionistic type theory.

[97] The first detailed analysis of the notion of presupposition was given by Duns Scotus, 'De rerum principio'.

[98] These two questions are the type-theoretic equivalents of what Quine calls *soundness* and *completeness* for a system of logic ('A proof procedure for quantification theory', p. 145).

(1) Is every demonstrable assertion correct?
(2) Is every correct assertion demonstrable?

That the answer to (1) is vehemently *yes* follows from what demonstration and correctness mean: if you reason according to valid inference rules you arrive at knowledge of the conclusion, whence the conclusion is correct. That is, intuitionistic type theory is *sound* since its inference rules are made evident. The answer to (2) depends on whether we are confronted with a complete or an incomplete assertion.

In the second case, the answer is *no*, according to Gödel's incompleteness theorem, if we consider the inference rules to be fixed.[99] If we allow new inference rules to be justified and added to the logical system, the answer becomes *yes in principle*.[100] *In principle* because, as indicated by Leibniz's principle of sufficient reason, in many cases, the causes cannot, practically speaking, be known to us.

For complete forms of assertion, the answer is again *yes in principle* if we allow new inference rules to be justified and added to the logical system. Thus, if *demonstrable* is taken to mean demonstrable by any valid inference rules, not fixed in advance, then *demonstrable* and *correct* coincide.

1.5 The Laws of Logic

As mentioned above, all standard laws of propositional logic, except the law of excluded middle, can be justified under the intuitionistic interpretation of the logical connectives.

The inference rule

$$\frac{A \text{ true} \quad B \text{ true}}{A \,\&\, B \text{ true}}$$

is self-evident upon remembering that a cause of $A \,\&\, B$ consists of a cause of A and a cause of B. Note that this inference rule is *well-formed* since the presupposition of the conclusion, that $A \,\&\, B$ is a proposition, follows from the presuppositions of the premisses, i.e., that A and B are propositions. Similarly, the inference rules

$$\frac{A \,\&\, B \text{ true}}{A \text{ true}}$$

[99] Gödel, 'Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I'. More specifically, fixing a collection of forms of expression and their corresponding inference rules, containing the expressions and rules of arithmetic, there are arithmetic propositions which cannot be demonstrated using only inference rules from this collection, but which are demonstrable, and hence correct, using valid inference rules outside of the collection.

[100] Cf., Martin-Löf, 'On the meanings of the logical constants and the justifications of the logical laws', p. 37. Note that, unless we fix our inference rules, the answer to (2) cannot be *no*. To answer *no* we have to know an assertion to be correct but not demonstrable, but the only way to come to know that an assertion is correct is though a demonstration.

and

$$\frac{A \ \& \ B \text{ true}}{B \text{ true}}$$

are well-formed and self-evident.

Remember that a cause of $A \vee B$ consists of a cause of A or a cause of B together with information about which cause it is that is given. This explanation makes the inference rules

$$\frac{A \text{ true} \quad (B : \text{prop})}{A \vee B \text{ true}}$$

and

$$\frac{(A : \text{prop}) \quad B \text{ true}}{A \vee B \text{ true}}$$

self-evident. In these inference rules, the premiss which is needed only to make the inference rule well-formed, i.e., as a presupposition of the conclusion, is put in parentheses. The other logical laws involving disjunction are the Stoic mood *modus tollendo ponens*[101] and proof by dilemma, which are expressed by the inference rules

$$\frac{A \vee B \text{ true} \quad \sim A \text{ true}}{B \text{ true}}$$

and

$$\frac{A \vee B \text{ true} \quad A \supset C \text{ true} \quad B \supset C \text{ true}}{C \text{ true}}$$

respectively. In these, and similar, inference rules, the leftmost premiss is called the major premiss and the other premisses are called minor premisses. The Stoic mood *modus tollendo ponens* can be justified directly using the meanings of the terms involved; but, as it can also be reduced to more primitive inference rules, this justification is left to the reader at this point. In the disjunctive syllogism, or proof by dilemma, the propositions A and B are called the *horns* of the dilemma and $A \supset C$ and $B \supset C$ are the two lemmata after which this mood of demonstration is named. The justification of proof by dilemma goes as follows: to get a cause of C, first inspect the cause of $A \vee B$; if this consists of a cause of A, invoke the left lemma with this cause of A to get a cause of C; if the cause of $A \vee B$ consists of a cause of B, invoke the right lemma with this cause of B to get a cause of C; in both cases, C has a cause.

[101] *Modus tollendo ponens*: mood which by denying affirms.

For implication,[102] the most important inference rule is *modus ponendo ponens*:

$$\frac{A \supset B \text{ true} \quad A \text{ true}}{B \text{ true}} .$$

Recall that a cause of $A \supset B$ consists of a method that takes any cause of A into a cause of B; a cause of A is given by the second premiss; combining these ingredients and performing the method results in a cause of B, i.e., the inference rule is evident upon explaining the meanings of the terms involved.

The connective Λ is associated with the logical law *ex falso quodlibet*, i.e., the inference rule

$$\frac{\Lambda \text{ true} \quad (A : \text{prop})}{A \text{ true}} .$$

This inference rule is justified as follows: granted that Λ has a cause c, a cause of A has to be given for each of the possible forms of c; there are no possible forms of c, so there is no work to be done; thus A has a cause. A perhaps more transparent way of seeing that this inference rule is valid is to compare it to proof by dilemma and *modus ponendo ponens*. The proposition $A \vee B$ is a binary disjunction; a unary disjunction is naturally identified with a proposition A; a nullary disjunction is false and thus identified with Λ. Thus, the propositions $A \vee B$, A, and Λ are in a falling scale. The corresponding inference rules are

$$\frac{A \vee B \text{ true} \quad A \supset C \text{ true} \quad B \supset C \text{ true}}{C \text{ true}}$$

with two minor premisses,

$$\frac{A \text{ true} \quad A \supset C \text{ true}}{C \text{ true}}$$

with one minor premiss, i.e., *modus ponendo ponens* with the premisses reversed, and

$$\frac{\Lambda \text{ true}}{C \text{ true}}$$

with no minor premisses.

[102] We have chosen to take the inference rule *modus ponendo ponens* as meaning determining for implication. In doing so we are faithful to the natural formulation of the BHK interpretation of $A \supset B$, namely that a cause of $A \supset B$ consists of a method taking a cause of A into a cause of B. Another interpretation which, *prima facie*, seems equivalent but which, in fact, is not, is that a cause of $A \supset B$ consists of a cause of B provided that a cause of A is given: this is the interpretation given by Kolmogorov, 'Zur Deutung der intuitionistischen Logik', p. 59, with the only difference that his interpretation is formulated in terms of problems and solutions instead of in terms of propositions and causes.

Since negation is defined in terms of implication and *falsum*, there are strictly speaking no inference rules which pertain to negation; instead inference rules involving negation are special cases of other inference rules. For example, the principle of noncontradiction

$$\frac{A \text{ true} \quad {\sim}A \text{ true}}{\Lambda \text{ true}}$$

is a special case of *modus ponendo ponens*.

Two of the Stoic moods remain, namely, *modus tollendo tollens*

$$\frac{A \supset B \text{ true} \quad {\sim}B \text{ true}}{{\sim}A \text{ true}}$$

and *modus ponendo tollens*

$$\frac{{\sim}(A \mathbin{\&} B) \text{ true} \quad A \text{ true}}{{\sim}B \text{ true}\ .}$$

These inference rules can either be justified directly, or demonstrated in terms of more basic inference rules.

There is an important but subtle difference between demonstrating something from known, or accepted, premisses and demonstrating it from premisses which are merely assumed, contingently, as it were. Properly speaking, inferences are made only in the former case, where we pass from something we know to something we get to know. An example will make this clearer. First, think about the letters L, M, P, and F as having the following meanings

$$\begin{cases} \mathrm{L} = \text{to be a logician,} \\ \mathrm{M} = \text{to be a mathematician,} \\ \mathrm{P} = \text{to be a philosopher,} \\ \mathrm{F} = \text{to be interested in first principles.} \end{cases}$$

Let it moreover be accepted that a logician is a philosopher or a mathematician, and that a philosopher is interested in first principles, i.e.,

$$\begin{cases} \mathrm{L} \supset (\mathrm{P} \vee \mathrm{M}) \text{ true, and} \\ \mathrm{P} \supset \mathrm{F} \text{ true}\ . \end{cases}$$

To get an example of demonstration properly speaking, think about somebody who is a logician but not interested in first principles, i.e., grant that L is true and that ∼F is true. It can now be demonstrated that the person you have in mind is in fact a mathematician:

$$\dfrac{\dfrac{\mathrm{L} \supset (\mathrm{P} \vee \mathrm{M})\ \text{true} \quad \mathrm{L}\ \text{true}}{\mathrm{P} \vee \mathrm{M}\ \text{true}} \quad \dfrac{\mathrm{P} \supset \mathrm{F}\ \text{true} \quad \sim\mathrm{F}\ \text{true}}{\sim\mathrm{P}\ \text{true}}}{\mathrm{M}\ \text{true}}\ .$$

On the other hand, if you do not have any particular person in mind but want to demonstrate the proposition

$$(\mathrm{L} \,\&\, \sim\mathrm{F}) \supset \mathrm{M},$$

i.e., that *if* somebody is a logician but not interested in first principles *then* he is a mathematician, then demonstration from merely assumed premisses has to be involved.

From Aristotle to Gentzen, logicians took for granted that demonstration from merely assumed premisses follows the same laws as demonstration from accepted premisses.[103] It could have been objected that this practice was unfounded, but we know of no such objection prior to Gentzen. Instead, Gentzen showed how demonstration from assumed premisses is to be understood in terms of demonstration from accepted premisses, and solved the problem at the same time as he formulated it.

When demonstrating propositions from assumed premisses the kind of propositions dealt with are *hypothetical*; in this context we understand any proposition of the form $A \supset B$ as hypothetical. Traditionally, the Stoic moods were called hypothetical syllogisms and their major premisses were all called hypothetical propositions, i.e., the propositions $A \vee B$ and $\sim(A \,\&\, B)$ were considered hypothetical, in addition to $A \supset B$[104]; with our definition of negation, the negated conjunction is hypothetical, but the disjunction is not.

The difference, mentioned above, between demonstrating something from accepted premisses and demonstrating something from assumed premisses can now be reformulated as follows: what is the difference between the *validity* of the inference rule

$$\dfrac{A_1\ \text{true} \quad \cdots \quad A_n\ \text{true}}{C\ \text{true}}$$

and the *truth* of the implication

$$(A_1 \,\&\, \cdots \,\&\, A_n) \supset C?$$

That an inference rule is valid means that, once the premisses are known, nothing more is called for to come to know the conclusion. That the implication is true

[103] Cf., Aristotle, *An. Pr.*, Bk. 1, Ch. 1; *An. Post.*, Bk. 1, Ch. 2; Gentzen, 'Untersuchungen über das logische Schließen I & II'; and Sundholm, 'Inference versus Consequence'.

[104] Cf., Boëthius, 'De hypotheticis syllogismis'.

means that there is a method which takes a cause of $A_1 \ \& \ \cdots \ \& \ A_n$ into a cause of C. In the inference rule *modus ponendo ponens*, a hypothetical proposition occurs as a premiss, so it seems as if inference is more fundamental than implication. That it has to be so is seen most clearly by Carroll's paradox.[105] If the validity of the inference rule

$$\frac{A \supset B \text{ true} \quad A \text{ true}}{B \text{ true}}$$

was dependent on the truth of the proposition

$$((A \supset B) \ \& \ A) \supset B,$$

we would need the inference rule

$$\frac{((A \supset B) \ \& \ A) \supset B \text{ true} \quad A \supset B \text{ true} \quad A \text{ true}}{B \text{ true}}$$

to reach the conclusion B, but then the validity of this inference rule would be dependent on the truth of the proposition

$$((((A \supset B) \ \& \ A) \supset B) \ \& \ (A \supset B) \ \& \ A) \supset B,$$

etc., *ad infinitum*. The conclusion that B is true would never be reached, as the poor Achilles experienced in Carroll's paradox.

It would be a terrible blow to logic if its laws could not be justified also in the hypothetical case, but, indeed, they can be.[106] We adopt natural deduction style notation

$$\frac{\begin{array}{c}[A \text{ true}] \\ \vdots \\ B \text{ true}\end{array}}{A \supset B \text{ true}}$$

when inferring the truth of an implication from a hypothetical demonstration of the truth of the consequent from the truth of the antecedent.

[105] Carroll, 'What the Tortoise said to Achilles'.

[106] As demonstrated by Gentzen, 'Untersuchungen über das logische Schließen I & II'. Cf., Granström, *Treatise on Intuitionistic Type Theory*, Ch. II, § 7.

1.6 The Intuitionistic Interpretation of Apagoge

To deny the equivalence of the propositions A and $\sim\sim A$ is a bold but necessary step to take: however, not all is lost. First, A entails $\sim\sim A$, as demonstrated by

$$\dfrac{\dfrac{A \text{ true} \quad [\sim A \text{ true}]^{\star}}{\Lambda \text{ true}}}{\sim\sim A \text{ true}\,.}\ {}^{\star}$$

Moreover, the two propositions $\sim A$ and $\sim\sim\sim A$ are equivalent.[107] One side of this equivalence is a special case of the law established above (with $\sim A$ for A), and the other side of the equivalence is demonstrated by

$$\dfrac{\sim\sim\sim A \text{ true} \quad \dfrac{[A \text{ true}]^{\star}}{\sim\sim A \text{ true}}}{\dfrac{\Lambda \text{ true}}{\sim A \text{ true}\,.}}\ {}^{\star}$$

Thus, negative propositions are equivalent to their double negation, but positive propositions need not be. Instead of *duplex negatio affirmat*, intuitionistic logic has *triplex negatio negat*.

Keeping these logical laws in mind, we will now investigate the distinction between the two assertions

$$A \text{ true}$$

and

$$\sim\sim A \text{ true}$$

in greater detail.

A distinction made by Aristotle in connection with syllogistic reasoning is between *direct* proof and *indirect* proof (proof *per impossibile*).[108] A direct proof proceeds by inference rules, as we are used to. In an indirect proof of A, one assumes the negation of A and shows that this assumption leads to a contradiction: with the intuitionistic interpretation of negation, this leads to an intuitionistic proof of $\sim\sim A$. The distinction between direct and indirect proofs was upheld by Kant, using the Greek words ostensive and apagogical.

[107] This was first demonstrated by Brouwer, 'Intuitionistische Zerlegung mathematischer Grundbegriffe', p. 253.

[108] Cf., Aristotle, *An. Pr.*, Bk. 2, Ch. 14.

> The third rule peculiar to pure reason, in so far as it is to be subjected to a discipline in respect of transcendental proofs, is that its proofs must never be *apagogical*, but always *ostensive*. The direct or ostensive proof, in every kind of knowledge, is that which combines with the conviction of its truth insight into the sources of its truth; the apagogical proof, on the other hand, while it can indeed yield certainty, cannot enable us to comprehend truth in its connection with the grounds of its possibility. The latter is therefore to be regarded rather as a last resort than as a mode of procedure which satisfies all the requirements of reason.[109]

In the history of logic, there is also another topic of importance to the distinction between A being true and $\sim\sim A$ being true, namely, the topic of causal proofs.[110] In brief: Aristotle made a distinction between demonstration of a fact (ὅτι) and demonstration of the reason for it (διότι).[111] In Latin, these terms were rendered *quia* and *propter quid*, i.e., demonstration *that* and demonstration *because of something*. Next, Averroes developed this distinction further by adding a third kind of demonstration, *potissima*, i.e., best of all, which is a simultaneous demonstration of the fact and the reason for it.[112] This distinction is called for if one admits inductive (or, better, abductive) reasoning from effect to cause, which then would be *propter quid* but not of a fact, because the conclusion is not necessary. Such demonstrations are not accepted in mathematics, whence we will make no further use of this distinction, but instead consider *propter quid* and *potissima* as synonymous. During the Renaissance, some authors claimed that there are no causes in mathematics, so its demonstrations cannot be *potissima*[113]; Biancani, among others, replied that the demonstrations of mathematics are *potissima* since they are by formal or material cause.[114] Indirect proofs were generally not considered causal.[115] Now the distinction became that between proofs that proceed by causes (*potissima*) and proofs that do not (*quia*): the former yield evidence while the latter only yield certainty. That is, something is certain if it cannot be otherwise and evident if known by its causes:

> Archimedes' admirers need to excuse his oblique procedure; both because it is long and complicated in the constructions and the proofs and because it is not completely satisfactory, since it produces certainty but not evidence. I am of the opinion that everything evident is certain but not everything certain is evident.[116]

[109] Kant, *Kritik der reinen Vernunft*, Pt. 2.1.4, p. 513 (B 817) (trans. N. K. Smith).

[110] For a comprehensive treatment of this topic, the reader is referred to the first two chapters of Mancosu's book *Philosophy of Mathematics and Mathematical Practice in the Seventeenth Century*.

[111] Aristotle, *An. Post.*, Bk. 1, Ch. 13; and ibid., Bk. 2, Ch. 1.

[112] Mancosu, *Philosophy of Mathematics and Mathematical Practice in the Seventeenth Century*, p. 12.

[113] Ibid., p. 13.

[114] Ibid., p. 17.

[115] Ibid., p. 25.

[116] Nardi, quoted in ibid., p. 63.

It is natural to identify a proposition A being *evident* in Nardi's sense with it being *true* in our sense, and a proposition being *certain* with $\sim\sim A$ being true; because A implies $\sim\sim A$ but not the other way around, i.e., "everything evident is certain but not everything certain is evident". Through this long and, admittedly, inconclusive line of argument, Aristotle's distinction between *quia* and *propter quid* is reduced to that between $\sim\sim A$ being true and A being true.

The use of the word *certain* is somewhat unfortunate in this context because it suggests some kind of epistemic modality. It is clear that the use of the word *certain* in Nardi's distinction between kinds of evidence established by different mathematical proofs or constructions due to Archimedes is not the same as the use of the word in the distinction between knowledge and certainty.[117] Rather, certainty in Nardi's sense is a kind of knowledge—but of what? We have already identified the sentence "A is evident" with the sentence "I know A is true" and the sentence "A is certain" with the sentence "I know $\sim\sim A$ is true"; but instead of saying that $\sim\sim A$ is true, we can say that A is *irrefutable*,[118] i.e., that its negation does not admit a proof. Now *true* and *irrefutable* are, as it were, on the same level: we can know that A is true and we can know that A is irrefutable.

In his introduction to intuitionism, Heyting makes use of the distinction between negation *de jure* and negation *de facto*[119]: the former is the intuitionistic negation, while the latter negation has the property that $\sim\sim A$ entails A. This distinction becomes clearer if we identify *de jure* negation with the negation of the proposition A in the assertion that A is evident, or true, and *de facto* negation with the negation of A in the assertion that A is irrefutable[120]: with this distinction, both negations are the ordinary intuitionistic negation, but if A is negated twice in the assertion that A is irrefutable, we get that $\sim\sim A$ is irrefutable, or, which amounts to the same, that $\sim\sim\sim\sim A$ is true, which entails that A is irrefutable. Thus, the terms *de jure* and *de facto* could instead be applied to the proposition A, just as evident and certain, i.e., that A *de facto* is true, or that A is a fact, can be taken to mean that $\sim\sim A$ is true, i.e., that A is irrefutable.

Finally, Bolzano revived the Aristotelian distinction between *quia* and *propter quid* and made a distinction between Gewissmachungen and Begründungen, i.e., certifications and groundings.[121] For Bolzano, this distinction is not the same as that between apagogical and ostensive, but, again, a lot of what is said about the difference between certifications and groundings makes sense when a certification is taken to be a demonstration of $\sim\sim A$ being true and a grounding a demonstration of A being true.

[117] As discussed in Moore's 1941 Howison lecture 'Certainty' and Wittgenstein's book *On Certainty*.

[118] This terminology was suggested by Sundholm (personal communication).

[119] Heyting, *Intuitionism: An Introduction*, p. 18.

[120] Cf., ibid., Th. 1, p. 17.

[121] Sebestik, 'Bolzano's Logic'.

Thus, the essence of the observations which lead the various authors to make these distinctions really is that between $\sim\sim A$ being true and A being true.

Let

$$A \text{ irrefutable}$$

be an abbreviation for $\sim\sim A$ true. That is, the bidirectional inference rule

$$\frac{\sim\sim A \text{ true}}{A \text{ irrefutable}}$$

is valid. As demonstrated above, every true proposition is irrefutable, i.e., we have the inference rule

$$\frac{A \text{ true}}{A \text{ irrefutable}\ .}$$

Moreover, irrefutability and truth coincide for negative propositions, i.e., we have the bidirectional inference rule

$$\frac{\sim A \text{ irrefutable}}{\sim A \text{ true}\ .}$$

Observe also that irrefutability and truth coincide for *falsum*,

$$\frac{\Lambda \text{ irrefutable}}{\Lambda \text{ true}\ .}$$

Intuitionistic logic is primarily concerned with what is true, i.e., evident or *per causas*. Fortunately, the laws of logic are valid also when dealing with apagogical knowledge, or knowledge of irrefutable propositions. In fact, all logical laws demonstrated above are valid with 'true' replaced by 'irrefutable'. This is just an alternative formulation of the double negation interpretation, first presented by Kolmogorov.[122]

Additional tools are available when demonstrating an irrefutable conclusion. The principle of proof by contradiction,[123] can be formulated as the following special case of implication introduction

[122] Kolmogorov, 'On the principle of excluded middle'. Cf., Glivenko, 'Sur quelques points de la logique de M. Brouwer'; Gödel, 'Zur intuitionistische Arithmetik und Zahlentheorie'; and Gentzen, 'Die Widerspruchfreiheit der reinen Zahlentheorie'. A direct (as opposed to metamathematical) demonstration of the logical laws for irrefutable propositions is given by the Author in *Treatise on Intuitionistic Type Theory*, Ch. VI, § 1.

[123] Also called proof *per contradictionem* or *per impossibile*, *reductio ad absurdum* or *ad impossibile*.

$$\frac{\begin{array}{c}[\sim A \text{ true}]\\ \vdots\\ \Lambda \text{ true}\end{array}}{A \text{ irrefutable .}}$$

That the conclusion of this inference rule is that A is irrefutable fits well with the view that proofs *per impossibile* do not give causal knowledge of the conclusion.[124]

A well-known result in intuitionistic logic is that every proposition of the form $A \vee \sim A$ is irrefutable.[125] To demonstrate this logical law, note that the inference rules

$$\frac{\sim(A \vee B) \text{ true}}{\sim A \text{ true}} \quad \text{and} \quad \frac{\sim(A \vee B) \text{ true}}{\sim B \text{ true}}$$

are both valid. The double negative form of the law of excluded middle is now demonstrated by

$$\frac{\dfrac{[\sim(A \vee \sim A) \text{ true}]^{\star}}{\sim\sim A \text{ true}} \quad \dfrac{[\sim(A \vee \sim A) \text{ true}]^{\star}}{\sim A \text{ true}}}{\dfrac{\Lambda \text{ true}}{A \vee \sim A \text{ irrefutable .}}\ ^{\star}}$$

If we combine the irrefutability of $A \vee \sim A$ with proof by dilemma we get the hypothetical inference rule

$$\frac{\begin{array}{cc}[A \text{ true}] & [\sim A \text{ true}]\\ \vdots & \vdots\\ B \text{ irrefutable} & B \text{ irrefutable}\end{array}}{B \text{ irrefutable,}}$$

which may be termed proof by cases.

Thus, for propositional logic, the classical laws of logic can be recovered by dealing with irrefutability instead of truth. Unfortunately, the same idea does not quite work for predicate logic.

It was recognised by the authors of *Principia Mathematica* that using the law of excluded middle (or its equivalent, proof by contradiction) to *prove* the law of excluded middle involves a vicious circle.[126] In view of this, it is astonishing that the critics of Brouwer's rejection of the law of excluded middle claimed that his

[124] Mancosu, *Philosophy of Mathematics and Mathematical Practice in the Seventeenth Century*, p. 25.

[125] Though not explicitly stated in this form, this insight is due to Brouwer, 'The Unreliability of the Logical Principles', p. 110.

[126] Whitehead et al., *Principia Mathematica*, Intro., Ch. 2, § 1, p. 40.

rejection leads to a third truth value, which is inconsistent,[127] and that Church had to correct his fellow logicians by restating that their argument involves a vicious circle.[128]

As for the third truth value, which allegedly is a consequence of denying the law of excluded middle, it might well be that the antagonists of intuitionism are referring to the state of doubt. With respect to knowledge, a man's attitude towards a proposition can be broadly divided into three: he may know the proposition to be true, he may know the proposition to be false, and he may know neither that it is true nor that it is false. Thus, true, doubtful, and false, are not three truth values, but, as it were, three knowledge states.

Logicians make a distinction between the law of excluded middle and the principle of bivalence. The law of excluded middle is usually formulated as the proposition $A \vee \sim A$ being true whenever A is a proposition. It is natural to equate this law with the (invalid) inference rule

$$\frac{A : \text{prop}}{A \vee \sim A \text{ true}}$$

in intuitionistic type theory. The principle of bivalence cannot be formulated as an inference rule in intuitionistic type theory—it has to be formulated in the metalanguage: for any proposition A, either A is true *or* $\sim A$ is true, with a metalinguistic *or*.

The validity of the principle of bivalence of course depends on the meaning assigned to the notions of proposition, truth, and negation, and the exact sense in which exactly one of A and $\sim A$ must be true; the validity of the law of excluded middle further depends on the meaning assigned to disjunction. Under the bivalent truth value interpretation of the notions involved,[129] both principles are valid.

In the very beginning of *Outlines of Pyrrhonism*, Sextus Empiricus makes the following observation:

> The natural result of any investigation is that the investigators either discover the object of search or deny that that it is discoverable and confess it to be inapprehensible or persist in their search. So, too, with regard to the objects investigated by philosophy, this is probably why some have claimed to have discovered the truth, others have asserted that it cannot be apprehended, while others again go on inquiring.[130]

Sextus Empiricus calls these three views dogmatic, academic, and sceptic, respectively—Sextus Empiricus himself of course being a sceptic. The three

[127] Cf., Mancosu, *From Brouwer to Hilbert*, pp. 278–280.

[128] Church, 'On the law of excluded middle', p. 77.

[129] According to which a proposition is interpreted as a truth value, i.e., as an element of the set {true, false}; the truth of a proposition A is interpreted as A being equal to 'true'; and negation and disjunction have their usual Boolean definitions.

[130] Empiricus, *Outlines of Pyrrhonism*, Bk. 1, Ch. 1.

possible outcomes of the search for an object are, in particular, applicable to the search for a cause of the truth of a proposition, and correspond to the three knowledge states mentioned above.

I will take the principle of bivalence to be tantamount to the principle that all doubt is possible to overcome: *non ignorabimus* to speak with Hilbert.[131] Here are some possible attitudes towards this principle.

(1) The most optimistic position is that there is a *systematic* method to establish either A true or $\sim A$ true, for any proposition A. I take this position to imply a positive solution to Hilbert's Entscheidungsproblem, in direct contradiction with the result gained by Church and Turing.[132] Thus, this position is self-contradictory.
(2) The second most optimistic position is to claim to know a (nonsystematic) method to establish either A true or $\sim A$ true, for any proposition A. Somebody in this position claims to have evidence for the law of excluded middle. This certainly entails the principle of bivalence since, if the intuitionistic disjunction $A \vee B$ is true, then A is true or B is true. I will call anybody in possession of a method for deciding any proposition an oracle.[133] Claiming to be an oracle seems both pathological and irrefutable.
(3) A third possibility is to claim that there is a method to establish either A true or $\sim A$ true, for any proposition A, without claiming to know such a method, i.e., to claim that oracles exist, without claiming to be one.
(4) A fourth position is that there *may* be a method to establish either A true or $\sim A$ true, for any proposition A, but that this method is not humanly attainable, i.e., the content of this position is that there are no human oracles.
(5) A fifth and less optimistic position is that there is no method which, for any proposition A, establishes either A true or $\sim A$ true, i.e., that there cannot be any oracles.
(6) Finally, the least optimistic position is that there is a proposition A for which it can be known to be impossible to establish A true and equally impossible to establish $\sim A$ true. This position is self-contradictory if we agree that we may infer that $\sim A$ is true from knowledge of the impossibility of establishing that A is true.[134] This entailment is reasonable since to know that it is impossible to establish A, one has to possess a method of producing an absurd consequence from an alleged cause of A, and this method is a cause of $\sim A$. So, for the alleged counterexample A to the principle of bivalence, we have A being both false and irrefutable, which is absurd.

[131] Cf., Hilbert, 'Mathematical problems', p. 445.

[132] Cf., Church, 'An unsolvable problem of elementary number theory' and Turing, 'On Computable Numbers'.

[133] The use of the word *oracle* in this connection was introduced by Turing, 'Systems of logic based on ordinals', § 4, p. 172.

[134] Cf., Martin-Löf, 'Verificationism Then and Now', Third Law, p. 16.

The connection between oracles and the principle of bivalence brings us to another classical topic, namely περὶ δυνατῶν, about things possible.[135] To establish the connection between the principle of bivalence and oracles, it suffices to apply the six positions on the principle of bivalence, discussed above, to propositions about the future.

According to Cicero, the ancients argued that if something was without cause, this would contradict the principle that every proposition was necessarily either true or false.[136] In our terminology, that something, A, is without cause can be interpreted as A being irrefutable without having a cause; this cannot happen if the principle of bivalence holds, because then $\sim A$ must have a cause if A does not, in contradiction to the assumption that A was irrefutable. Thus, the principle of bivalence implies that no fact, i.e., irrefutable proposition, is without cause. Cicero reports that, from this implication, Chrysippus argued, by *modus ponendo ponens*, that all things take place by fate, and Epicurus, by *modus tollendo tollens*, that not every proposition is necessarily either true or false:

> At this point, in the first place if I chose to agree with Epicurus and to say that not every proposition is either true or false, I would rather suffer that nasty knock than agree that all events are caused by fate; for the former opinion has something to be said for it, but the latter is intolerable.[137]

That Chrysippus' position is intolerable shows that to avoid fatalism, we have to deny the principle of bivalence, i.e., we have to take position five above, at least if we take the notion of proposition in the most general possible sense, including propositions about the future, and take the principle of bivalence to mean that either A is true *now* or $\sim A$ is true *now*.[138]

This does not settle the question whether the principle of bivalence holds for that which is actual or for that which is timeless, like mathematics. Aristotle escapes the problem by making this distinction:

> For one half of the said contradiction must be true and other half false. But we cannot say which half is which. Though it may be that one is more probable, it cannot be true yet or false. There is evidently, then, no necessity that one should be true, the other false, in the case of affirmations and denials. For the case of those things which as yet are potential, not actually existent, is different from that of things actual.[139]

This can be read as a denial of the most general form of the principle of bivalence, while maintaining that it holds for propositions about the present, i.e., about things actual.

---

[135] Cicero, *De Fato*, Ch. 1.

[136] Ibid., Ch. 10, beginning.

[137] Ibid., Ch. 10, n. 21.

[138] This is the most natural interpretation of the principle of bivalence, since the assertion A true can be expanded into *I know a logical cause of A*, in which the *now* is implicit.

[139] Aristotle, *Perih.*, Ch. 9, 19a37–19b5.

To maintain the principle of bivalence for actual propositions, i.e., that every proposition about the present is either true or false, entails that every proposition about the future *will become* either true or false. If we accept this principle we have to beware of an error which is easy to make, viz., to claim that if two persons hold contradictory propositions about the future, one of them is right and the other wrong. It is not so, because to know is to know by causes, and, most likely, both of them are wrong, i.e., speaking without knowing. Put differently, if you make a guess, and it turns out as you predicted, your guess was still not knowledge, i.e., you did not speak the truth. This kind of reasoning seems to have confused Cicero:

> For it is necessary that of two contradictory propositions, *pace* Epicurus, that one should be true and the other false; for example, 'Philoctetes will be wounded' was true, and 'Philoctetes will not be wounded' false, for the whole of the ages of the past; unless perhaps we choose to follow the opinion of the Epicureans, who say that propositions of this sort are neither true nor false, or else, when ashamed of that, they nevertheless make the still more impudent assertion that disjunctions consisting of contradictory propositions are true, but that the statements contained in the propositions are neither of them true. What marvellous effrontery and pitiable ignorance of logical method![140]

It is interesting to note that the position of the intuitionists agrees rather well with that of the Epicureans, as reported by Cicero: they deny the law of excluded middle, i.e., the truth of the proposition $A \vee \sim A$, and, when ashamed of that, affirm the irrefutability of the proposition $A \vee \sim A$, and deny the principle of bivalence.

A final objection to the principle of bivalence and the law of excluded middle, this time even for propositions about the present and the timeless, is that it fails to hold because of an intrinsic vagueness in the terms involved in the proposition at hand.[141] Problems of this kind are related to the old paradoxes about the bald man and the heap[142]: how many hairs may a man have and still be called bald? how many stones make a heap? If, for every number n, the proposition *n stones make a heap* is either true or false, there must be a least number for which it is true, contrary to intuition. To get the unintuitive conclusion, we have to use the law of excluded middle. An often overlooked virtue of intuitionism is that it dissolves this kind of paradoxes: we can affirm that one or two stones do not make a heap and that fifty or more stones make a heap without having to make up our minds for the numbers in between.

Acknowledgements I would like to express my great appreciation to Prof. Shahid Rahman for suggesting and encouraging the composition of this summary of the Author's book *Treatise on Intuitionistic Type Theory*, and to Springer for permission to reuse parts of the book. The book version of this paper was written under supervision of Prof. Per Martin-Löf, and I would like to offer him my special thanks for his signficant time investment and his constant precision of thought and expression.

[140]Cicero, *De Fato*, Ch. 16, nn. 37–38. I have changed the translation to conform with standard terminology in logic by replacing the word contrary with the word contradictory and removing Cicero's comment giving an explanation of his unusual sense of the word contrary.

[141]Cf., Geach, 'The law of excluded middle', pp. 71–73.

[142]Kneale et al., *The Development of Logic*, p. 114.

References

Aquinas, S.T.: Summa Theologiae. In: Opera omnia. Iussu impensaque Leonis XIII P. M. edita, vols. 4–12. Ex typographia polyglotta S. congregationis de propaganda fide, Rome (1888–1906)

Aquinas, S.T., Cathala, M.-R., Spiazzi, R.M. (eds.): In duodecim libros Metaphysicorum Aristotelis expositio. Marietti, Turin (1950)

Aquinas, S.T.: De Veritate. In: Spiazzi, R. (ed.) Quaestiones Disputatae, vol. 1. Marietti, Turin (1953)

Aquinas, S.T.: In Peri Hermeneias. In: In Aristotelis Libros Peri Hermeneias et Posteriorum Analyticorum Expositio, 2nd edn. Marietti, Turin (1964)

Aquinas, S.T.: In Posteriorum Analyticorum. In: In Aristotelis Libros Peri Hermeneias et Posteriorum Analyticorum Expositio, 2nd edn. Marietti, Turin (1964)

Aquinas, S.T.: De ente et essentia. In: Opera omnia. Iussu impensaque Leonis XIII P. M. edita, vol. 43, pp. 315–381. Editori di San Tommaso, Rome (1976)

Aristotle: Metaphysics (Trans. by H. Tredennick). Loeb Classical Library, vol. 271/287. Harvard University Press, Cambridge (1933/1935)

Aristotle: Analytica priora (Trans. by H. Tredennick). Loeb Classical Library, vol. 325, pp. 181–531. Harvard University Press, Cambridge (1938)

Aristotle: Perihermenias (Trans. by H.P. Cooke). Loeb Classical Library, vol. 325, pp. 111–179. Harvard University Press, Cambridge (1938)

Aristotle: Analytica posteriora (Trans. by H. Tredennick). Loeb Classical Library, vol. 391, pp. 1–261. Harvard University Press, Cambridge (1960)

Aristotle: Topica (Trans. by E.S. Forster). Loeb Classical Library, vol. 391, pp. 263–739. Harvard University Press, Cambridge (1960)

Aristotle: The Physics (Trans. by P.H. Wicksteed and F.M. Cornford). Loeb Classical Library, vols. 228 and 255. Harvard University Press, Cambridge (1986/2006)

Arnauld, A.: The Art of Thinking (Trans. by J. Dickoff and P. James). Bobbs-Merrill, New York (1964)

Bacon, F.: Novum Organum. J Billius, London (1620)

Becker, O.: Mathematische Existenz. Zur Logik und Ontologie mathematischer Phänomene. Jahrb. für Philos. und phänomenologische Forsch. **8**, 441–809 (1927)

Bernays, P.: Mathematische Existenz und Widerspruchsfreiheit. Etudes de Philosophie des sciences en hommage á Ferdinand Gonseth, pp. 11–25. Griffon, Neuchatel (1950)

Bishop, E., Bridges, D.: Constructive Analysis. Springer, Berlin (1985)

Blancanus, J.: A treatise on the nature of mathematics along with a chronology of outstanding mathematicians. In: Mancosu, P. (ed.) Philosophy of Mathematics and Mathematical Practice in the Seventeenth Century (Trans. by G. Klima), pp. 178–212. Oxford University Press, Oxford (1996)

Bocheński, I.M.: Ancient Formal Logic. North-Holland, Amsterdam (1951)

Boëthius, S.: De syllogismo categorico. In: Migne, J.-P. (ed.) Patrologiae cursus completus. Series Latina, vol. 64, pp. 793C–832A. Garnier, Paris (1847)

Boëthius, S.: De hypotheticis syllogismis. In: Obertello, L. (ed.) Logicalia, vol. 1. Brescia, Parma (1968)

Bolzano, B.: Wissenschaftslehre. Seidel, Sulzbach (1837)

Boole, G.: The calculus of logic. Camb. Dublin Math. J. **3**, 183–198 (1848)

Bourbaki, N.: Elements of Mathematics. Theory of Sets. Addison-Wesley, Boston (1968)

Brouwer, L.E.J.: Intuitionistische Zerlegung mathematischer Grundbegriffe. Jahresber. d. Deutsch. Math. Vereinig. **33**, 251–256 (1925)

Brouwer, L.E.J.: The unreliability of the logical principles. In: Heyting, A. (ed.) Collected Works. Philosophy and Foundations of Mathematics, vol. 1, pp. 107–111. North Holland, Amsterdam (1975) (original from 1908)

Buridan, J.: Summulae de Dialectica (Trans. by G. Klima). Yale University Press, New Haven (2001)
Cajetan, T.: Commentaria in Praedicamenta Aristotelis. Angelicum, Rome (1939)
Cajori, F.: A History of Mathematics. Macmillan, New York (1919)
Cajori, F.: Mathematical signs of equality. Isis **5**(1), 116–125 (1923)
Carroll, L.: What the Tortoise said to Achilles. Mind **4**(14), 278–280 (1895)
Church, A.: On the law of excluded middle. Bull. Am. Math. Soc. **34**, 75–78 (1928)
Church, A.: An unsolvable problem of elementary number theory. Am. J. Math. **58**(2), 345–363 (1936)
Cicero, M.T.: De Fato (Trans. by H. Rackham). Loeb Classical Library. Harvard University Press, Cambridge (1942)
Cocchiarella, N.B.: Conceptual realism as a formal ontology. In: Poli, R., Simons, P. (eds.) Formal Ontology, pp. 27–60. Kluwer, Dordrecht (1996)
Coffey, P.: The Science of Logic, 2nd edn., vol. 1. Peter Smith, New York (1938)
De Morgan, A.: Formal Logic. Taylor and Walton, London (1847)
Descartes, R.: The world or treatise on light. In: The Philosophical Writings of Descartes (Trans. by R. Stoothoff). Cambridge University Press, Cambridge (1985)
Di Bella, S.: Causa Sive Ratio. Univocity of reason and plurality of causes in Leibniz. In: Dascal, M. (ed.) Leibniz: What Kind of Rationalist? Logic, Epistemology, and the Unity of Science, vol. 13. Springer, Dordrecht (2008). Chap. 32
Diogenes, L.: Lives of Eminent Philosophers (Trans. by R.D. Hicks). Loeb Classical Library, vols. 1/2. Harvard University Press, Cambridge (1925)
Empiricus, S.: Outlines of Pyrrhonism (Trans. by R.G. Bury). Loeb Classical Library, vol. 1. Harvard University Press, Cambridge (1961)
Frege, G.: Begriffsschrift. Louis Nebert, Halle (1879)
Frege, G.: Grundgesetze der Arithmetik II, vol. 2. Hermann Pohle, Jena (1903)
Frege, G.: On concept and object. In: Translations from the Philosophical Writings of Gottlob Frege (Trans. by P.T. Geach), pp. 42–55. B Blackwell, Oxford (1960)
Geach, P.T.: The law of excluded middle. Supp. Proc. Arist. Soc. **30**, 59–90 (1956)
Geach, P.T.: Identity. Rev. Metaphys. **21**(1), 3–12 (1967)
Geach, P.T.: Logic Matters. Blackwell, Oxford (1972)
Gentzen, G.: Untersuchungen über das logische Schließen I & II. Math. Zeit. **39**, 176–210 & 405–431 (1935)
Gentzen, G.: Die Widerspruchfreiheit der reinen Zahlentheorie. Math. Ann. **112**, 493–565 (1936)
Glivenko, V.: Sur quelques points de la logique de M. Brouwer. Acad. R. de Belg. Bull. **15**, 183–188 (1929)
Gödel, K.: Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I. Monatshefte für Math. und Phys. **38**, 173–198 (1931)
Gödel, K.: Zur intuitionistische Arithmetik und Zahlentheorie. Ergeb. eines math. Kolloqu. **4**, 34–38 (1933)
Granström, J.G.: Treatise on Intuitionistic Type Theory. Logic, Epistemology and the Unity of Science. Springer, Dordrecht (2011)
Gredt, I.: Elementa Philosophiae Aristotelico-Thomisticae, 4th edn., vol. 1. Herder, Freiburg (1925)
Heyting, A.: Die formalen Regeln der intuitionistischen Logik. Sitzungsberiche der Preussischen Akademie der Wissenschaften, 42–56 (1930)
Heyting, A.: Sur la logique intuitionniste. Acad. R. de Belg. Bull. **16**, 957–963 (1930)
Heyting, A.: Intuitionism: An Introduction. North-Holland, Amsterdam (1956)
Hilbert, D.: Mathematical problems (Trans. by M.W. Newson). Bull. Am. Math. Soc. **8**(10), 437–479 (1902)
Hilbert, D.: On the infinite. In: van Heijenoort, J. (ed.) From Frege to Gödel. A Source Book in Mathematical Logic, 1879–1931, pp. 367–392. Harvard University Press, Cambridge (1967)

Hilbert, D., Bernays, P.: Grundlagen der Mathematik, vol. 1. Springer, Berlin (1934)
Husserl, E.: Logische Untersuchungen, 3rd edn., vol. 1. M Niemeyer, Halle (1922)
Husserl, E.: Logische Untersuchungen, 5th edn., vol. 2. M Niemeyer, Tübingen (1968)
Husserl, E.: The Crisis of European Sciences and Transcendental Phenomenology (Trans. by D. Carr). Northwestern University Press, Evanston (1970)
Kant, I.: Kritik der reinen Vernunft. Kants Werke, vol. 3. de Gruyter, Berlin (1968)
Kleene, S.C.: On the interpretation of intuitionistic number theory. J. Symb. Log. **10**(4), 109–124 (1945)
Klev, A.: Categories and logical syntax. PhD thesis, Leiden University (2014)
Kneale, W., Kneale, M.: The Development of Logic. Oxford University Press, Oxford (1962)
Kolmogorov, A.: Zur Deutung der intuitionistischen Logik. Math. Zeit. **35**, 58–65 (1932)
Kolmogorov, A.: On the principle of excluded middle. In: From Frege to Gödel. A Source Book in Mathematical Logic, 1879–1931 (Original in Russian from 1925), pp. 416–437. Harvard University Press, Cambridge (1967)
Leibniz, G.W.: Principia Philosophiæ. Petrus Conrad Monath, Frankfurt (1728)
Locke, J.: An Essay Concerning Humane Understanding, 2nd edn. Thomas Basset, London (1694)
Maddy, P.: Mathematical existence. Bull. Symb. Log. **11**(3), 351–376 (2005)
Mancosu, P.: Philosophy of Mathematics and Mathematical Practice in the Seventeenth Century. Oxford University Press, Oxford (1996)
Mancosu, P. (ed.): From Brouwer to Hilbert. Oxford University Press, New York (1998)
Maritain, J.: The Degrees of Knowledge (Trans. by R. McInerny). University of Notre Dame Press, Notre Dame (1995)
Markov, A.A.: The theory of algorithms. Am. Math. Soc. Transl. Ser. 2 **15**, 1–14 (1960)
Martin-Löf, P.: An intuitionistic theory of types: predicative part. In: Rose, H.E., Shepherdson, J. (eds.) Logic Colloquium '73, pp. 73–118. North-Holland, Amsterdam (1975)
Martin-Löf, P.: About models for intuitionistic type theories and the notion of definitional equality. In: Proceedings of the Third Scandinavian Logic Symposium. Studies in Logic and the Foundation of Mathematics, vol. 82, pp. 81–109. North-Holland, Amsterdam (1975)
Martin-Löf, P.: Intuitionistic Type Theory. Studies in Proof Theory. Bibliopolis, Napoli (1984)
Martin-Löf, P.: A path from logic to metaphysics. In: Atti del Congresso Nuovi Problemi della Logica e della Filosofia Scienza, Viareggio, 1990, vol. 2, pp. 141–149. Clueb, Bologna (1991)
Martin-Löf, P.: Analytic and synthetic judgements in type theory. In: Parrini, P. (ed.) Kant and Contemporary Epistemology, pp. 87–99. Kluwer (1994)
Martin-Löf, P.: Verificationism then and now. In: De Pauli-Schimanovich, W., Köhler, E., Stadler, F. (eds.) The Foundational Debate; Complexity and Constructivity in Mathematics and Physics, pp. 187–196. Kluwer (1995)
Martin-Löf, P.: On the meanings of the logical constants and the justifications of the logical laws. Nord. J. Philos. Log. **1**(1), 11–60 (1996)
Mill, J.S.: A System of Logic. Harper & Brothers, New York (1846)
Moore, G.E.: The nature of judgement. Mind **8**(30), 176–193 (1899)
Moore, G.E.: Certainty. In: Philosophical Papers. Allen and Unwin, London (1959)
Nordström, B., Petersson, K., Smith, J.M.: Programming in Martin-Löf's Type Theory. Oxford University Press (1990)
Peano, G.: Arithmetices Principia Nova Methodo Exposita. Fratelli Bocca, Turin (1889)
Poinsot, J.: The Material Logic of John of St. Thomas. Basic Treatises (Trans. by Y.R. Simon, J.J. Glanville, and G.D. Hollenhorst). The University of Chicago Press, Chicago (1955)
Quine, W.V.: A proof procedure for quantification theory. J. Symb. Log. **20**(2), 141–149 (1955)
Quine, W.V.: Theories and Things. Harvard University Press, Cambridge (1981)
Recorde, R.: The Whetstone of Witte. Da Capo Press, Amsterdam (1969)
Russell, B.: The theory of implication. Am. J. Math. **28**(2), 159–202 (1906)
Russell, B.: Mathematical logic as based on the theory of types. Am. J. Math. **30**(3), 222–262 (1908)

Scotus, J.D.: De rerum principio. In: Hibernicus, M. (ed.) Quaestiones Subtilissime Scoti in Meta-physicam Aristotelis. Eiusdem de Primo Rerum Principio Tractatus Atque Theoremata. Bonetus Locatellus, Venice (1497)
Sebestik, J.: Bolzano's logic. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy. The Metaphysics Research Lab, Stanford (2007)
Simpson, S.G.: Logic and mathematics. In: Rosen, S. (ed.) The Examined Life, Readings from Western Philosophy from Plato to Kant, pp. 577–605. Random House, New York (2000)
Skolem, T.A.: Some remarks on axiomatized set theory. In: van Heijenoort, J. (ed.) From Frege to Gödel. A Source Book in Mathematical Logic, 1879–1931, pp. 290–301. Harvard University Press, Cambridge (1967)
Sundholm, G.: Existence, proof, and truth-making: a perspective on the intuitionistic conception of truth. Topoi **13**, 117–126 (1994)
Sundholm, G.: Inference versus consequence. In: Childers, T. (ed.) The LOGICA Yearbook 1997, pp. 26–35. Filosofia, Prague (1998)
Turing, A.M.: On computable numbers, with an application to the entscheidungsproblem. Proc. Lond. Math. Soc. **2**(42), 230–265 (1936)
Turing, A.M.: Systems of logic based on ordinals. Proc. Lond. Math. Soc. Ser. 2 **45**(2239), 161–228 (1939)
Virgil: Georgics (Trans. by H.R. Fairclough), 2nd edn. Loeb Classical Library. Harvard University Press, Cambridge (1999)
Weyl, H.: Über die neue Grundlagenkrise der Mathematik. Math. Zeit. **10**, 39–79 (1921)
Weyl, H.: Comments on Hilbert's second lecture on the foundations of mathematics. In: van Heijenoort, J. (ed.) From Frege to Gödel. A Source Book in Mathematical Logic, 1879–1931, pp. 480–484. Harvard University Press, Cambridge (1967)
Whitehead, A.N., Russell, B.: Principia Mathematica, vol. 1. Cambridge University Press, Cambridge (1910)
Wittgenstein, L.: Tractatus Logico-Philosophicus (Trans. by C.K. Ogden and F.P. Ramsey). Routledge & Kegan Paul, London (1922)
Wittgenstein, L., Anscombe, G.E.M., von Wright, G.H.: On Certainty (Trans. by D. Paul and G.E.M. Anscombe). Basil Blackwell, Oxford (1969)

Chapter 2
Atomic Systems in Proof-Theoretic Semantics: Two Approaches

Thomas Piecha and Peter Schroeder-Heister

Abstract Atomic systems are systems of rules containing only atomic formulas. In proof-theoretic semantics for minimal and intuitionistic logic they are used as the base case in an inductive definition of validity. We compare two different approaches to atomic systems. The first approach is compatible with an interpretation of atomic systems as representations of states of knowledge. The second takes atomic systems to be definitions of atomic formulas. The two views lead to different notions of derivability for atomic formulas, and consequently to different notions of proof-theoretic validity. In the first approach, validity is stable in the sense that for atomic formulas logical consequence and derivability coincide for any given atomic system. In the second approach this is not the case. This indicates that atomic systems as definitions, which determine the meaning of atomic sentences, might not be the proper basis for proof-theoretic validity, or conversely, that standard notions of proof-theoretic validity are not appropriate for definitional rule systems.

Keywords Proof-theoretic semantics • Atomic systems • Higher-level rules • Definitions • Definitional reflection • Minimal logic • Intuitionistic logic

2.1 Introduction

Within proof-theoretic semantics for logical constants the validity of atomic formulas, or atoms, is usually defined in terms of derivability of these formulas in atomic systems. Such systems can be sets of atomic formulas, figuring as atomic axioms, or sets of atomic rules, that is, of rules which only contain atomic formulas. Examples of such rules are production rules or definite Horn clauses. One can also allow for atomic rules which can discharge atomic assumptions, or even consider higher-level atomic rules which can discharge assumed atomic rules. Further crucial use of atomic systems is made in explaining the logical constant of implication. An implication $A \to B$ is valid with respect to an atomic system S (in short: S-valid)

T. Piecha (✉) • P. Schroeder-Heister
Department of Computer Science, University of Tübingen, Sand 13, 72076 Tübingen, Germany
e-mail: thomas.piecha@uni-tuebingen.de; psh@uni-tuebingen.de

J. Redmond et al. (eds.), *Epistemology, Knowledge and the Impact of Interaction*,
Logic, Epistemology, and the Unity of Science 38, DOI 10.1007/978-3-319-26506-3_2

if and only if *for all extensions S' of S* it holds that whenever A is S'-valid then B is S'-valid. The reference to extensions guarantees that validity is monotone with respect to atomic systems. Otherwise it might happen that a formula, which is valid with respect to S, is invalid with respect to an extension of S.

This monotonicity requirement is motivated by the interpretation of atomic systems as knowledge bases. What is valid should remain valid, if our knowledge as incorporated in an atomic knowledge base is extended. However, there are contexts in which we do not expect monotonicity to hold, as studied, for example, in the various branches of non-monotonic logic. Here we study definitional contexts as a particular case. When we interpret atomic systems as definitions, we cannot postulate monotonicity. If we extend the definition of a term, a valid proposition may lose its validity. Correspondingly, when basing proof-theoretic validity on atomic systems as definitions, for the S-validity of implication and consequence we should not refer to arbitrary extensions S' of S. In his recent publications, Prawitz, who coined the notion of proof-theoretic validity in Prawitz (1971), prefers this definitional reading of atomic systems ('bases') and explicitly refrains from the reference to extensions of atomic systems:

> A base is seen as determining the meanings of the atomic sentences. (Prawitz 2016, section 5)
>
> To consider extensions of the given base [...] is natural when a base is seen as representing a state of knowledge, but is in conflict with the view adopted here that a base is to be understood as giving the meanings of the atomic sentences. (Prawitz 2016, fn. 12)

This view leads to problems, however. We will show that, if validity is based on atomic systems understood as definitions, then it is not *stable*, that is, logical consequence and derivability diverge already at the atomic level. This negative result is even independent of whether consequence and implication are characterized with respect to arbitrary extensions of atomic systems or not. This shows that the definitional view of atomic systems is not compatible with the concept of proof-theoretic validity in its given form. This result depends, of course, on the theory of definitions used. In this paper we rely on the approach based on the idea of definitional reflection (see Hallnäs 1991, 2006; cf. Schroeder-Heister 1993) according to which the definitional reading of atomic rules is implemented by a rule schema which expresses that the clauses given for a certain atom *exhaustively* characterize that atom.

We confine ourselves to propositional logic, as this suffices to make our point. In Sect. 2.2 we consider notions of proof-theoretic validity which are monotone with respect to extensions of atomic systems. In Sect. 2.3 we compare this approach to Kripke semantics and show that proof-theoretic validity corresponds to considering validity in a specific Kripke model. In Sect. 2.4 we describe the idea of atomic systems as definitions and establish that stability is lost under the definitional reading of atomic systems. Derivability from assumptions and validity of consequence do not even coincide in the atomic case.

2.2 Atomic Systems and Proof-Theoretic Validity

2.2.1 First-Level Atomic Systems and Validity

Atomic systems have been considered in proof-theoretic approaches to validity by Prawitz (1971) and Dummett (1991), for example. There atomic systems are sets of production rules for atomic formulas, or atoms, $a, b, \ldots, a_1, a_2, \ldots$, defined as follows:

Definition 2.1. A *(first-level) atomic system* S is a (possibly empty) set of atomic rules of the form

$$\frac{a_1 \quad \ldots \quad a_n}{b}$$

where the a_i and b are atoms. The set of premisses $\{a_1, \ldots, a_n\}$ in a rule can be empty; in this case the rule is an *atomic axiom* and of *level 0*.

The *derivability* of an atom a from a (possibly empty) set $\{a_1, \ldots, a_n\}$ of atomic assumptions in an atomic system S is written $a_1, \ldots, a_n \vdash_S a$. *Derivations* are defined as usual. For example, for the atomic system S:

$$\frac{}{c} \qquad \frac{a}{b} \qquad \frac{b \quad c}{d}$$

the derivation

$$\dfrac{\dfrac{a}{b} \quad \dfrac{}{c}}{d}$$

shows $a \vdash_S d$.

Extensions S' of atomic systems S are understood in the set-theoretic sense, that is, an atomic system S' is an *extension* of an atomic system S, written $S' \supseteq S$, if S' results from adding a (possibly empty) set of atomic rules to S. For example, $S' = S \cup \{a\}$ is an extension of S by the atomic axiom a. For this extension $\vdash_{S'} d$ holds.

In proof-theoretic notions of validity, the validity of atoms is determined by their derivability in atomic systems, and the validity of complex formulas is defined inductively with respect to such systems. Originally, Prawitz (1971, 1973, 1974, 2014) gave certain notions of validity for derivations which are constructed from arbitrary inference rules. These notions of validity not only depend on atomic systems but also on reduction procedures ('justifications') which transform such derivations into other derivations (see also Schroeder-Heister 2006, 2012).

In what follows, we consider instead notions of validity for formulas (see Piecha et al. 2014), which do not depend on reduction procedures. We restrict ourselves

to formulas $A, B, \ldots$ in the fragment $\{\to, \vee, \wedge\}$ of minimal propositional logic; absurdity $\bot$ is just a distinguished atom, not a logical constant.

Definition 2.2. *S-validity* ($\models_S$) and *validity* ($\models$) are defined as follows:

(S1) $\models_S a \ :\Longleftrightarrow\ \vdash_S a$,
(S2) $\models_S A \to B \ :\Longleftrightarrow\ A \models_S B$,
(S3) $\Gamma \models_S A \ :\Longleftrightarrow\ \forall S' \supseteq S : (\models_{S'} \Gamma \implies \models_{S'} A)$, where Γ is a set of formulas, and where $\models_{S'} \Gamma$ stands for $\{\models_{S'} A_i \mid A_i \in \Gamma\}$,
(S4) $\models_S A \vee B \ :\Longleftrightarrow\ \models_S A$ or $\models_S B$,
(S5) $\models_S A \wedge B \ :\Longleftrightarrow\ \models_S A$ and $\models_S B$,
(S6) $\Gamma \models A \ :\Longleftrightarrow\ \forall S : \Gamma \models_S A$.

By clause (S1), S-validity of atoms is defined in terms of derivability in an atomic system S. Another important use of atomic systems is made in the definition of S-consequence $\Gamma \models_S A$ (S3), and thus of S-validity of implication $\models_S A \to B$ (S2), which is defined by S-consequence $A \models_S B$. In clause (S3), arbitrary extensions of atomic systems are considered. This has the effect that an S-consequence $\Gamma \models_S A$ cannot just hold because some atom on which Γ depends is not valid in S. This would be the case if S-consequence $\Gamma \models_S A$ were, for example, defined by

$$\Gamma \models_S A \ :\Longleftrightarrow\ (\models_S \Gamma \implies \models_S A) \tag{S3$'$}$$

where no extensions of S are considered. In this case, if, for example, $\Gamma = \{a\}, A = b$ and $S = \emptyset$, then $\not\models_S a$ and thus trivially $(\models_S a \implies \models_S b)$, and hence $a \models_S b$. Validity with respect to atomic systems would therefore fail to be monotone, since for example for $S' = S \cup \{a\} = \{a\}$ we have $a \not\models_{S'} b$ while $a \models_S b$. This situation is avoided by considering arbitrary extensions in the definition of S-consequence. Indeed, taking extensions into account guarantees monotonicity, as we can easily prove:

$$\Gamma \models_S A \implies \forall S' \supseteq S : \Gamma \models_{S'} A.$$

2.2.2 Higher-Level Atomic Systems

Atomic systems need not be restricted to systems of first level. Second-level and arbitrary higher-level atomic systems can be considered as well (see Piecha et al. 2014; cf. Schroeder-Heister 1984; Sandqvist 2015).

Definition 2.3. A *second-level atomic system* S is a (possibly empty) set of atomic rules of the form

$$\frac{\begin{array}{ccc}[\Gamma_1] & & [\Gamma_n] \\ a_1 & \ldots & a_n\end{array}}{b}$$

where the a_i and b are atoms, and the Γ_i are finite sets of atoms. The sets Γ_i may be empty, in which case the rule is a *first-level rule*. The set of premisses $\{a_1, \ldots, a_n\}$ can be empty as well; in this case the rule is an axiom.

Such a rule can be applied as follows: If the premisses $a_1, \ldots, a_n$ have been derived in S from certain assumptions $\Gamma_1, \ldots, \Gamma_n$, then one may conclude b, where, for each i, in the branch of the subderivation leading to a_i assumptions belonging to Γ_i may be discharged.

Second-level atomic systems are now further generalized to the higher-level case by allowing for atomic rules which can discharge not only atoms but atomic rules as assumptions (see Schroeder-Heister 1984, 2014; Olkhovikov and Schroeder-Heister 2014; cf. Piecha et al. 2014).

Definition 2.4. We use the following linear notation for atomic *higher-level rules*:

1. Every atom a is a rule of level 0.
2. If $R_1, \ldots, R_n$ are rules ($n \geq 1$), whose maximal level is ℓ, and a is an atom, then $(R_1, \ldots, R_n \rhd a)$ is a rule of level $\ell + 1$.

In tree notation, higher-level rules have the form

$$\frac{\begin{array}{ccc}[\Gamma_1] & & [\Gamma_n] \\ a_1 & \ldots & a_n\end{array}}{b}$$

where the a_i and b are atoms, and the Γ_i are finite sets $\{R_1^i, \ldots, R_k^i\}$ of rules, which may be empty. The set of premisses $\{a_1, \ldots, a_n\}$ of such a rule can again be empty, in which case the rule is an axiom.

Definition 2.5. A *higher-level atomic system* S is a (possibly empty) set of higher-level rules.

Higher-level rules can be represented by formulas in the fragment $\{\rightarrow, \wedge\}$:

Definition 2.6. With every rule R in a set of rules S a formula R^* representing R is associated as follows:

1. $a^* := a$, for atoms a.
2. $(R_1, \ldots, R_n \rhd a)^* := R_1^* \wedge \ldots \wedge R_n^* \rightarrow a$, for a rule $R_1, \ldots, R_n \rhd a$.

Then S^* is defined as the set of formulas representing the rules in S.

In the higher-level case, atomic *rules* can be used as (dischargable) assumptions, whereas in the second-level case only atoms could be used in that way. This difference requires a definition of the notion of *derivation* of atoms from rules:

Definition 2.7. For a level-0 rule a,

$$\frac{}{a}\, a$$

is a *derivation* of a from $\{a\}$.

Now consider a level-$(\ell+1)$ rule $(\Gamma_1 \rhd a_1), \ldots, (\Gamma_n \rhd a_n) \rhd b$. Suppose that for each i $(1 \leq i \leq n)$ a derivation

$$\begin{array}{c} \Sigma_i \cup \Gamma_i \\ \mathscr{D}_i \\ a_i \end{array}$$

of a_i from $\Sigma_i \cup \Gamma_i$ is given. Then

$$\dfrac{\begin{array}{ccc} \Sigma_1 & & \Sigma_n \\ \mathscr{D}_1 & & \mathscr{D}_n \\ a_1 & \ldots & a_n \end{array}}{b}\,(\Gamma_1 \rhd a_1), \ldots, (\Gamma_n \rhd a_n) \rhd b$$

is a *derivation* of b from $\Sigma_1 \cup \ldots \cup \Sigma_n \cup \{(\Gamma_1 \rhd a_1), \ldots, (\Gamma_n \rhd a_n) \rhd b\}$.

An atom b is *derivable* from Σ in a higher-level atomic system S, symbolically $\Sigma \vdash_S b$, if there is a derivation of b from $\Sigma \cup S$.

We give an example derivation for the atomic system

$$S \left\{ \begin{array}{l} (b \rhd e) \rhd f \\ ((a \rhd b) \rhd c) \rhd e \end{array} \right.$$

and the set of assumptions $\Sigma = \{((a \rhd b) \rhd d), ((d, b) \rhd c)\}$:

$$\dfrac{\dfrac{\dfrac{\dfrac{\overline{a}\,[a]^1}{b}\,[a \rhd b]^2}{d}\,1\,(a \rhd b) \rhd d \qquad \overline{b}\,[b]^3}{c}\,(d, b) \rhd c}{e}\,2\,\langle((a \rhd b) \rhd c) \rhd e\rangle}{f}\,3\,\langle(b \rhd e) \rhd f\rangle$$

The derivation shows $\Sigma \vdash_S f$. (Angle brackets $\langle\,\rangle$ are used to indicate the rules of S, and square brackets [ ] with numerals indicate the discharge of assumptions.)

The definition of validity for second-level or higher-level atomic systems is exactly the same as that for first-level atomic systems (Definition 2.2). The generalization from first- to higher-level atomic systems does not affect the monotonicity of validity: S-validity, and hence validity, for higher-level atomic systems is monotone with respect to extensions $S' \supseteq S$.

2.2.3 Completeness Issues

It can be shown that minimal logic is not complete with respect to validity. A counterexample is the consequence

$$a \to (b \vee c) \vDash (a \to b) \vee (a \to c)$$

which holds independently of the level of atomic systems. Here it is important that a, b and c are individual atoms, not propositional variables. This counterexample ceases to hold for arbitrary substitutions of complex formulas for atoms. If, for example, $b \vee c$ is substituted for a, then the resulting consequence is no longer valid. This shows that validity is not closed under substitution. Since derivability in minimal logic is closed under substitution, one could demand that a notion of validity proposed for minimal logic should be closed under substitution as well. This can be done by definition:

Definition 2.8. *S-validity under substitution* ($\Vdash_S$) and *validity under substitution* ($\Vdash$) are defined as follows:

1. $\Gamma \Vdash_S A \ :\Longleftrightarrow$ for each substitution instance Γ', A' of $\Gamma, A : \Gamma' \vDash_S A'$.
2. $\Gamma \Vdash A \ :\Longleftrightarrow$ for each substitution instance Γ', A' of $\Gamma, A : \Gamma' \vDash A'$.

These strengthened notions of validity can be extended to intuitionistic logic. There one considers the following notion of validity:

Definition 2.9. Let $(\bot)$ stand for the set of rules $\left\{\dfrac{\bot}{a} \;\middle|\; a \text{ atomic}\right\}$. Then *intuitionistic S-validity* is defined as follows: $\Gamma \vDash^i_S A \ :\Longleftrightarrow \ \Gamma \vDash_{S\cup(\bot)} A$.

Intuitionistic validity $\Gamma \vDash^i A$ is defined as $\Gamma \vDash_{(\bot)} A$, and the corresponding notions closed under substitution, $\Gamma \Vdash^i_S A$ and $\Gamma \Vdash^i A$, are defined as $\Gamma \Vdash_{S\cup(\bot)} A$ and $\Gamma \Vdash_{(\bot)} A$, respectively.

For the case of higher-level atomic systems S it could be shown (see Piecha et al. 2014) that intuitionistic propositional logic is not complete for intuitionistic validity under substitution ($\Vdash^i$). A counterexample is the intuitionistically non-derivable but valid Harrop formula (where $\neg A := A \to \bot$):

$$(\neg A \to (B \vee C)) \to ((\neg A \to B) \vee (\neg A \to C)).$$

If we restrict ourselves to first-level atomic systems, the question of completeness is still open. However, in view of the fact that proof-theoretic validity amounts to considering a single Kripke model rather than the totality of all Kripke models (see Sect. 2.3 below), we would conjecture that, as in the higher-level case, we lose the completeness of intuitionistic logic. Proof-theoretic validity characterizes at best (that is, if validity is closed under substitution) some intermediate logic between the intuitionistic and classical systems.

For details concerning completeness we refer to Piecha et al. (2014) and Piecha (2016). Here we just remark that completeness (or failure of completeness) of logical systems for the proposed notions of validity depends essentially on the kind of atomic systems on which these notions are based.

2.2.4 Stability of S-Validity

Let Δ^* be the set of formulas representing a finite set Δ of atomic rules (in the sense of Definition 2.6). One can show that S-validity is *stable* in the sense that

$$\Delta^* \models_S b \iff \Delta^* \vdash_S b$$

holds for any atomic systems S. This includes *atomic completeness*

$$a_1, \ldots, a_n \models_S b \implies a_1, \ldots, a_n \vdash_S b$$

and *atomic soundness*

$$a_1, \ldots, a_n \vdash_S b \implies a_1, \ldots, a_n \models_S b$$

as special cases. Intuitionistic S-validity ($\models^i_S$) is stable as well.

Stability is an important feature of S-validity, since it guarantees that S-validity is not creative in the sense that atomic completeness fails, and that it is not destructive in the sense that atomic soundness fails (see also the discussion in Sandqvist (2015) on conservativeness as a desideratum). If we consider notions of S-validity which lack stability, we must take into account that atomic derivability from assumptions $a_1, \ldots, a_n \vdash_S a$ can be different from the corresponding S-consequence, even though atomic derivability $\vdash_S a$ is (by definition) equivalent with the S-validity of a. In this case, atomic systems would be used merely as a device to generate valid atoms, where the induced relation of derivability from assumptions can be totally disregarded. Technically, this is no problem. However, conceptually, this would not be much different from looking at atomic systems as sets of atoms which are valid by definition.

2.3 Proof-Theoretic Validity and Kripke Semantics

The formulation of Definition 2.2 has a striking resemblance to the definition of validity in Kripke semantics (see e.g. Troelstra 1988; van Dalen 2013; Moschovakis 2014). It can actually be viewed as a definition of validity in a special Kripke model.

In Kripke semantics for propositional intuitionistic logic a Kripke model $\mathscr{K}$ consists of a partial order $\leq$ between objects called *nodes* (or *reference points* or *worlds*) together with a valuation function v which tells which atoms are true at which node. Thus $v(a, k) = 1$ means that the atom a is true at node k. This valuation function must satisfy the monotonicity condition that, if $k' \geq k$ and $v(a, k) = 1$, then $v(a, k') = 1$. Intuitively, this means that what is true at some stage, must remain true. Then the validity $\models^{\mathscr{K}}_k A$ of a formula A in $\mathscr{K}$ at a node k, the validity $\Gamma \models^{\mathscr{K}}_k A$ of a consequence of A from Γ in $\mathscr{K}$ at k, the validity $\Gamma \models^{\mathscr{K}} A$ of a consequence of A

from Γ in $\mathscr{K}$, and the validity (simpliciter) $\Gamma \models A$ of a consequence of A from Γ (i.e., logical validity) are defined as follows:

Definition 2.10.

(K1) $\models_k^{\mathscr{K}} a \ :\Longleftrightarrow\ v(a,k) = 1$,
(K2) $\models_k^{\mathscr{K}} A \to B \ :\Longleftrightarrow\ A \models_k^{\mathscr{K}} B$,
(K3) $\Gamma \models_k^{\mathscr{K}} A \ :\Longleftrightarrow\ \forall k' \geq k : (\models_{k'}^{\mathscr{K}} \Gamma \implies \models_{k'}^{\mathscr{K}} A)$, where Γ is a set of formulas, and where $\models_{k'}^{\mathscr{K}} \Gamma$ stands for $\{\models_{k'}^{\mathscr{K}} A_i \mid A_i \in \Gamma\}$,
(K4) $\models_k^{\mathscr{K}} A \vee B \ :\Longleftrightarrow\ \models_k^{\mathscr{K}} A$ or $\models_k^{\mathscr{K}} B$,
(K5) $\models_k^{\mathscr{K}} A \wedge B \ :\Longleftrightarrow\ \models_k^{\mathscr{K}} A$ and $\models_k^{\mathscr{K}} B$,
(K6) $\Gamma \models^{\mathscr{K}} A \ :\Longleftrightarrow\ \forall k : \Gamma \models_k^{\mathscr{K}} A$,
(K7) $\Gamma \models A \ :\Longleftrightarrow\ \forall \mathscr{K} : \Gamma \models^{\mathscr{K}} A$.

As in Definition 2.2 we restrict ourselves to minimal logic. Normally, in Kripke semantics, the consequence relations $\Gamma \models_k^{\mathscr{K}} A$ and $\Gamma \models^{\mathscr{K}} A$ are not defined; instead, the validity of implication in $\mathscr{K}$ at node k is defined as:

$$\models_k^{\mathscr{K}} A \to B \ :\Longleftrightarrow\ \forall k' \geq k : (\models_{k'}^{\mathscr{K}} A \implies \models_{k'}^{\mathscr{K}} B).$$

However, it can easily be seen that our Definition 2.10 comes to the same, as far as the relations validity $\models_k^{\mathscr{K}} A$ of a formula and logical validity $\Gamma \models A$ are concerned.

From the parallelism between (S1)–(S6) and (K1)–(K6) it is obvious that the definition of validity in Definition 2.2 is the definition of validity for a *specific* Kripke model $\mathscr{S}$, the nodes of which are the atomic systems S, the accessibility relation $\leq$ between nodes is the inclusion relation $\subseteq$ between atomic systems, and the valuation function v is defined by the derivability in S, that is, $v(a,S) = 1 \ :\Longleftrightarrow\ \vdash_S a$. From this definition of v and the fact that $\leq$ is set inclusion $\subseteq$ it is clear that the monotonicity condition required for v is satisfied. $\Gamma \models A$ in the sense of Definition 2.2 means the same as $\Gamma \models^{\mathscr{S}} A$ for this Kripke model $\mathscr{S}$.

From this point of view the counterexamples to completeness mentioned in Sect. 2.2.3 and established in Sandqvist (2009), de Campos Sanz et al. (2014), Piecha et al. (2014) and Piecha (2016) are not really surprising. If the definition of validity is merely based on validity in a *specific* Kripke model, we cannot expect completeness for intuitionistic (here: minimal) logic, of which we know that it holds with respect to logical validity, that is, to validity in *all* Kripke models. There is no obvious reason why the model $\mathscr{S}$ should be 'canonical' in that it represents the totality of all Kripke models.

2.4 Atomic Systems as Definitions

If atomic systems are understood as knowledge bases, then monotonicity of validity with respect to extensions is certainly a desired property, since increased knowledge should at least account for what is already known. If a consequence $\Gamma \models_S A$ has

been established on the basis of some knowledge given by the atomic system S, and an atomic system S' extends that knowledge, then $\Gamma \models_{S'} A$ should hold as well.

There is, however, an alternative view of atomic systems, in which one would not expect monotonicity of consequence with respect to extensions. Atomic systems can be understood as definitions of atoms. As an extension of a definition changes in general what is being defined, it is to be expected that there are consequences which hold with respect to the initial definition but do no longer hold with respect to an extension of that definition.

2.4.1 Definitional Closure

Consider an atomic system

$$S \left\{ \begin{array}{l} \Gamma_1 \rhd a \\ \quad \vdots \\ \Gamma_k \rhd a \end{array} \right.$$

of k higher-level atomic rules. This can be read as a *definition* of the atom a by *defining conditions* Γ_i, for $1 \leq i \leq k$. In this definitional reading the atomic rules $\Gamma_i \rhd a$ are also called *definitional clauses*. The defining conditions in such clauses can be empty. In the terminology of inductive definitions (see Aczel 1977) one can thus distinguish basis clauses of the form $\emptyset \rhd a$ (or just a) and inductive clauses of the form $\Gamma_i \rhd a$ (for non-empty Γ_i).

A direct application of such a definition consists in passing from some defining condition Γ_i of a to the defined atom a:

$$\frac{\Gamma_i}{a}$$

Inferences of this kind are also called steps of *definitional closure*. They correspond to the individual steps in a derivation of an atom in a higher-level atomic system.

2.4.2 Definitional Reflection

In the reading of atomic systems as definitions a difference is introduced by the fact that in the case of a definition of an atom a it is assumed that nothing else defines a. This assumption, the extremality condition, is usually made only implicitly (for example in mathematical definitions), just by saying that something is a definition. Sometimes it is stated explicitly by saying that the clauses for a in a definition define *the smallest set* of objects for which the given clauses hold, or by adding a clause, the extremal clause, saying that *nothing else defines a*.

When this assumption is taken into consideration, an additional reasoning principle becomes available for definitions. For an atom a defined by

$$S\left\{\begin{array}{l}\Gamma_1 \rhd a\\ \quad\vdots\\ \Gamma_k \rhd a\end{array}\right.$$

one can, in addition to definitional closure, also reason by *definitional reflection* (see Hallnäs 1991, 2006; cf. Schroeder-Heister 1993):

$$\frac{a \qquad \begin{bmatrix}\Gamma_1\\ C\end{bmatrix} \quad \ldots \quad \begin{bmatrix}\Gamma_k\\ C\end{bmatrix}}{C}$$

This rule says that whenever a formula C follows from each of the defining conditions Γ_i of an atom a, then C follows from the defined atom a alone.

If no additional logical rules are available, or if no additional rules are available for the decomposition or construction of higher-level rules, then C will in general be an atomic formula. An exception is the case where a is an undefined atom, say $\bot$, that is, where S does not contain any clauses of the form $\Gamma \rhd \bot$. Then any formula C can be inferred from $\bot$ by definitional reflection, since the set of defining conditions of the undefined atom $\bot$ is empty. This means that for atomic systems S as definitions a principle of *ex falso quodlibet*

$$\bot \vdash_S C$$

is available as long as at least one atom $\bot$ is undefined in S.

Definitional reflection is only justified for atomic systems as definitions, that is, when an extremality condition is assumed. Without this assumption only definitional closure can be used.

2.4.3 Properties of Derivability

In general, a *definition* is any finite atomic system

$$S\left\{\begin{array}{lcl}\Gamma_1^1 \rhd a_1 & & \Gamma_1^n \rhd a_n\\ \quad\vdots & \cdots & \quad\vdots\\ \Gamma_{k_1}^1 \rhd a_1 & & \Gamma_{k_n}^n \rhd a_n\end{array}\right.$$

Definitions in this sense need not have basis clauses $\emptyset \rhd a_i$. They are thus similar to logic programs, where such a restriction is not made either.

We here consider only atomic systems of higher-level atomic rules, which could be represented by formulas in the fragment $\{\rightarrow, \wedge\}$ (see Definition 2.6). When atomic systems are used as definitions one could also allow the defining conditions $\Gamma^i_{j_i}$ in definitional clauses $\Gamma^i_{j_i} \rhd a_i$ to be arbitrary formulas (see Hallnäs and Schroeder-Heister 1990, 1991). However, this is not permitted in our setting here.

As an example, consider the following definition:

$$S \begin{cases} \Gamma \rhd a & \Gamma \rhd b \\ \Delta \rhd a & \Delta \rhd b \\ & \Sigma \rhd b \end{cases}$$

Using definitional closure and definitional reflection we can show that $a \vdash_S b$ (but not $b \vdash_S a$) holds:

$$\dfrac{\overset{1}{a} \qquad \dfrac{[\Gamma]^1}{b}\,(\text{def. closure}),\ \langle \Gamma \rhd b\rangle \qquad \dfrac{[\Delta]^1}{b}\,(\text{def. closure}),\ \langle \Delta \rhd b\rangle}{b}\,(\text{def. reflection on } S)\ {}^1$$

The set of subderivations $\left\{\dfrac{\Gamma}{b}, \dfrac{\Delta}{b}\right\}$ shows that b can be derived from each of the defining conditions of a, namely Γ and Δ. Thus definitional reflection can be applied to a, discharging the assumptions Γ and Δ. Without definitional reflection, $a \vdash_S b$ cannot be shown.

For the extension $S' = S \cup \{\Theta \rhd a\}$ we do not have $a \vdash_{S'} b$, if $\Theta \vdash_{S'} b$ does not hold. In other words, since b cannot be derived from each of the defining conditions of a (the exception being Θ), we cannot apply definitional reflection here, and it thus cannot be shown that b is derivable in S' from a as the only assumption. This example shows that atomic systems behave quite differently when they are treated as definitions. It shows in particular that derivability fails to be monotone with respect to extensions of atomic systems: For the given $S' \supseteq S$ we have $a \vdash_S b$ but $a \nvdash_{S'} b$. Monotonicity is already lost in the case of first-level atomic systems, as can be seen by letting the defining conditions $\Gamma, \Delta, \Sigma, \Theta$ be sets of atoms. By the same argument we can see that for the extension $S'' = S \cup \{a \rhd a\}$ we do not have $a \vdash_{S''} b$, because for that to hold we would already need $a \vdash_{S''} b$, which is exactly what we want to prove. In effect, the addition of the clause $a \rhd a$ to a definition blocks the application of definitional reflection with respect to a, as one of the premisses of definitional reflection would already require as proven what one intends to prove.

2.4.4 Validity Based on Definitions

We now consider S-validity and validity in the context of atomic systems as definitions. That is, we now consider the situation where derivability $\vdash_S$ is defined with respect to atomic systems S understood as definitions.

We distinguish two cases. In the first case S-validity and validity are exactly as given by Definition 2.2, where S-consequence $\Gamma \models_S A$ is defined using extensions $S' \supseteq S$:

$$\Gamma \models_S A \;:\Longleftrightarrow\; \forall S' \supseteq S : (\models_{S'} \Gamma \implies \models_{S'} A) \tag{S3}$$

In the second case we consider validity without extensions, that is, we define S-consequence $\Gamma \models_S A$ as follows:

$$\Gamma \models_S A \;:\Longleftrightarrow\; (\models_S \Gamma \implies \models_S A) \tag{S3$'$}$$

We show that atomic soundness fails for validity using extensions, and that atomic completeness fails for validity without extensions. In each case we give a very simple counterexample which only uses the framework of first-level rules.

Case 1: Validity with extensions. Atomic soundness does not hold. For the empty definition $S = \emptyset$ we have $a \vdash_S b$ by definitional reflection, since a is not defined. Now consider the extension $S' = S \cup \{a\} = \{a\}$ in which a is defined. Then $\vdash_{S'} a$ and thus $\models_{S'} a$, while $\nvdash_{S'} b$ and therefore $\not\models_{S'} b$. Hence $\forall S' \supseteq S : (\models_{S'} a \implies \models_{S'} b)$ fails to hold, which means $a \not\models_S b$.

Case 2: Validity without extensions. Atomic completeness does not hold. The definition $S = \{a \rhd a\}$ yields a counterexample. We have $\nvdash_S a$ and thus $\not\models_S a$; hence $a \models_S b$ by clause (S3$'$). But $a \nvdash_S b$, since in S only a can be derived from a.

Summing up, we have:

Proposition 2.1. *S-validity (with or without extensions) is not stable.*

As this result is independent of whether extensions are considered or not, it hints at a deeper issue in the relation between definitional bases and proof-theoretic validity. In definitional reasoning, consequence $\Gamma \vdash_S a$ is based on specific definitional rules, in particular on rules, which allow one to assume an atom in a specific way by means of definitional reflection. This has the effect that the biconditional

$$\Gamma \vdash_S a \iff (\vdash_S \Gamma \implies \vdash_S a)$$

is no longer guaranteed. On the other hand, proof-theoretic validity is fundamentally based on the biconditional

$$\Gamma \vDash_S a \iff (\vDash_S \Gamma \implies \vDash_S a)$$

(we disregard extensions). This suggests that definitional reasoning and proof-theoretic validity aim at different notions of consequence and therefore implication. It is possible indeed to build a notion of validity on top of definitional bases. However, this would not proceed according to a validity definition as set out in Definition 2.2, but by considering introduction and elimination rules for logical constants as instances of definitional rules and thus by incorporating logic into the realm of definitional reasoning (cf. de Campos Sanz and Piecha 2009; Schroeder-Heister 2016). Definitional reflection would then be considered to be a general reasoning principle which applies to the atomic and logical cases likewise.

2.5 Conclusion

We considered two approaches to atomic systems. They show that within proof-theoretic semantics widely differing notions of validity can be formulated, depending on how atomic systems are understood. The first approach dealt with atomic systems of production rules (first-level), of assumption-discharging rules (second-level) and of arbitrary higher-level rules, which allow for the discharge of assumed atomic rules. Such atomic systems can be understood as knowledge bases. Notions of proof-theoretic validity based on these kinds of atomic systems are monotone with respect to extensions of atomic systems. The choice of the kind of atomic systems can make a difference with respect to completeness (see Piecha et al. 2014; Piecha 2016).

In the second approach, where atomic systems are understood as definitions, the situation is quite different. The additional principle of definitional reflection induces a derivability relation which is not monotone with respect to extensions of such systems. It is doubtful whether notions of proof-theoretic validity in the sense of Definition 2.2 should be based on atomic systems understood as definitions: Atomic soundness does not hold for validity using extensions, and atomic completeness fails for validity not using extensions. This means that S-validity is not stable. Definitional reflection is a principle leading to a different notion of validity. Besides the points mentioned in the last paragraph of Sect. 2.4, definitional reflection goes beyond the scope of atomic systems, since in principle it allows one to derive not only atoms from atoms but also complex formulas from atoms. Although the underlying definitions are atomic systems, they might then no longer be foundational for the meaning explanations for the logical constants given in standard notions of proof-theoretic validity.

Acknowledgements This work was carried out within the French-German ANR-DFG project "Beyond Logic", DFG grant Schr 275/17-1. We thank an anonymous reviewer for helpful comments and suggestions.

References

Aczel, P.: An introduction to inductive definitions. In: Barwise, J. (ed.) Handbook of Mathematical Logic, pp. 739–782. North-Holland, Amsterdam (1977)

de Campos Sanz, W., Piecha, T.: Inversion by definitional reflection and the admissibility of logical rules. Rev. Symb. Log. **2**(3), 550–569 (2009)

de Campos Sanz, W., Piecha, T., Schroeder-Heister, P.: Constructive semantics, admissibility of rules and the validity of Peirce's law. Log. J. IGPL **22**(2), 297–308 (2014). First published online 6 Aug 2013

Dummett, M.: The Logical Basis of Metaphysics. Duckworth, London (1991)

Hallnäs, L.: Partial inductive definitions. Theor. Comput. Sci. **87**, 115–142 (1991)

Hallnäs, L.: On the proof-theoretic foundation of general definition theory. In: Kahle, R., Schroeder-Heister, P. (eds.) Proof-Theoretic Semantics. Synthese, vol. 148, pp. 589–602. Springer, Berlin (2006)

Hallnäs, L., Schroeder-Heister, P.: A proof-theoretic approach to logic programming. I. Clauses as rules. J. Log. Comput. **1**, 261–283 (1990)

Hallnäs, L., Schroeder-Heister, P.: A proof-theoretic approach to logic programming. II. Programs as definitions. J. Log. Comput. **1**, 635–660 (1991)

Moschovakis, J.: Intuitionistic logic. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy (2014). http://plato.stanford.edu/archives/fall2014/entries/logic-intuitionistic/

Olkhovikov, G.K., Schroeder-Heister, P.: Proof-theoretic harmony and the levels of rules: generalised non-flattening results. In: Moriconi, E., Tesconi, L. (eds.) Second Pisa Colloquium in Logic, Language and Epistemology, pp. 245–287. ETS, Pisa (2014)

Piecha, T.: Completeness in proof-theoretic semantics. In: Piecha, T., Schroeder-Heister, P. (eds.) Advances in Proof-Theoretic Semantics. Trends in Logic, vol. 43, pp. 231–251. Springer, Cham (2016)

Piecha, T., de Campos Sanz, W., Schroeder-Heister, P.: Failure of completeness in proof-theoretic semantics. J. Philos. Log. **44**(3), 321–335 (2015). First published online 1 Aug 2014

Prawitz, D.: Ideas and results in proof theory. In: Fenstad, J.E. (ed.) Proceedings of the Second Scandinavian Logic Symposium. Studies in Logic and the Foundations of Mathematics, vol. 63, pp. 235–307. North-Holland, Amsterdam (1971)

Prawitz, D.: Towards a foundation of a general proof theory. In: Suppes, P., et al. (eds.) Logic, Methodology and Philosophy of Science IV, pp. 225–250. North-Holland, Amsterdam (1973)

Prawitz, D.: On the idea of a general proof theory. Synthese **27**, 63–77 (1974)

Prawitz, D.: An approach to general proof theory and a conjecture of a kind of completeness of intuitionistic logic revisited. In: Pereira, L.C., Haeusler, E.H., de Paiva, V. (eds.) Advances in Natural Deduction. Trends in Logic, vol. 39, pp. 269–279. Springer, Berlin (2014)

Prawitz, D.: On the relation between Heyting's and Gentzen's approaches to meaning. In: Piecha, T., Schroeder-Heister, P. (eds.) Advances in Proof-Theoretic Semantics. Trends in Logic, vol. 43, pp. 5-25. Springer, Cham (2016)

Sandqvist, T.: Classical logic without bivalence. Analysis **69**, 211–217 (2009)

Sandqvist, T.: Base-extension semantics for intuitionistic sentential logic. Log. J. IGPL **22**(1), 147–154 (2015)

Sandqvist, T.: Hypothesis-discharging rules in atomic bases. In: Wansing, H. (ed.) Dag Prawitz on Proofs and Meaning. Outstanding Contributions to Logic, vol. 7, pp. 313–328. Springer, Cham (2015)

Schroeder-Heister, P.: A natural extension of natural deduction. J. Symb. Log. **49**, 1284–1300 (1984)

Schroeder-Heister, P.: Rules of definitional reflection. In: Proceedings of the Eighth Annual IEEE Symposium on Logic in Computer Science, Montreal 1993, pp. 222–232. IEEE Computer Society, Los Alamitos (1993)

Schroeder-Heister, P.: Validity concepts in proof-theoretic semantics. In: Kahle, R., Schroeder-Heister, P. (eds.) Proof-Theoretic Semantics. Synthese, vol. 148, pp. 525–571. Springer, Berlin (2006)

Schroeder-Heister, P.: Proof-theoretic semantics. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy (Winter 2012 Edition) (2012). http://plato.stanford.edu/archives/win2012/entries/proof-theoretic-semantics/

Schroeder-Heister, P.: The calculus of higher-level rules, propositional quantifiers, and the foundational approach to proof-theoretic harmony. Stud. Log. **102**, 1185–1216 (2014). Special issue, Gentzen's and Jaśkowski's Heritage: 80 Years of Natural Deduction and Sequent Calculi, edited by A. Indrzejczak

Schroeder-Heister, P.: Open problems in proof-theoretic semantics. In: Piecha, T., Schroeder-Heister, P. (eds.) Advances in Proof-Theoretic Semantics. Trends in Logic, vol. 43, pp. 253–283. Springer, Cham (2016)

Troelstra, A.S., van Dalen, D.: Constructivism in Mathematics: An Introduction, vol. 1. North-Holland, Amsterdam (1988)

van Dalen, D.: Logic and Structure, 5th edn. Springer, London (2013)

Chapter 3
Knowledge and Its Game-Theoretical Foundations: The Challenges of the Dialogical Approach to Constructive Type Theory

Shahid Rahman, Radmila Jovanovic, and Nicolas Clerbout

Abstract It is our main claim that the time is ripe to link the dynamic turn launched by game-theoretical approaches to meaning with P. Martin-Löf's Constructive Type Theory (CTT). Furthermore, we also claim that the dialogical framework provides the appropriate means to develop such a link. We will restrict our study to the discussion of two paradigmatic cases of dependences triggered by quantifiers, namely the case of the Axiom of Choice and the study of anaphora, that are by the way two of the most cherished examples of Hintikka.

Keywords Knowledge • Constructive type theory • Epistemic logic • Dialogical logic • Game-theoretical semantics

3.1 Introduction

Since the emergence of logic as a scientific discipline in the ancient tradition the interface between knowledge, reasoning and logic grew up as constituting a tight braid that structured the dynamics of public and scientific debates and more generally of rational argumentative interaction and decision making. However,

The present paper is part of an ongoing project in the context of the research-program "Argumentation, Decision, Action" (ADA) and the project LACTO both supported by the *Maison Européenne des Sciences de l' Homme et de la Société* – USR 318 and by the laboratory UMR 8163: STL.

S. Rahman (✉)
Université de Lille, UMR 8163: STL, Villeneuve-d'Ascq, France
e-mail: shahid.rahman@univ-lille3.fr

R. Jovanovic
University of Belgrade, Belgrade, Serbia
e-mail: jovanovic_radmila@yahoo.com

N. Clerbout
CDHACS-Universidad de Valparaíso, Valparaíso, Chile
e-mail: nicolas.clerbout@uv.cl

J. Redmond et al. (eds.), *Epistemology, Knowledge and the Impact of Interaction*,
Logic, Epistemology, and the Unity of Science 38, DOI 10.1007/978-3-319-26506-3_3

around the dawn of the twentieth century, the braid loosened and fell apart into separate threads. In fact, it is during the years that followed immediately after the failure of the logical positivism project that the links between science as a body of knowledge and the study of the process by which knowledge is achieved and grounded were cut off.[1]

Nevertheless, around 1960 epistemic approaches that echoed the old traditions, challenged the mainstream current that followed from the work of A. Tarski, K. Gödel and P. Bernays.[2] Those epistemic approaches, which, a while later, were called, following Michael Dummett, *antirealists*, found their formal argument in the mathematics of Brouwer and intuitionistic logic, while the others persisted with the formal background of the Tarski tradition, where Cantorian set theory is linked via model theory to classical logic. The point is that while for intuitionists the notion of proposition is based on a theory of meaning where knowledge plays a crucial role and where acquisition of such knowledge is expressed by a judgement, the model-theoretic tradition took truth rather than knowing the truth as their foundations of formal semantics. Intuitionists, as pointed out by D. Prawitz (2012, p. 47) *avoid the term truth and reject the idea that intuitionism could replace "p is true" with "there exists a proof of p" understood in a realistic vein*. Indeed, the existence of a proof, as pointed out by Prawitz in the same text, is to be *understood epistemically as the actual experience of the construction intended by the proposition, not as the existence of an ontological fact*. More generally, from the intuitionistic point of view proof theory provides the means for the development of an epistemic approach to meaning rooted in assertions (rather than propositions). In this context, it should be mentioned that already in 1955 Paul Lorenzen proposed an *operative* approach that delved into the conceptual and technical bonds between *procedure* and knowledge.[3] The insights of Lorenzen's *Operative Logik*, as pointed out by Schröder-Heister (2008), had lasting consequences in the literature on proof theory and still deserve attention nowadays. Indeed, the proof-theoretical notion of *harmony* formulated by the logicians that favoured the epistemic approach such as Dag Prawitz[4] has been influenced by Lorenzen's notions of *admissibility*, *eliminability* and *inversion*.[5]

However, the epistemic perspectives did not all reduce to the proof-theoretical framework: epistemic features were also implemented via game-theoretical approaches. Indeed, on one hand, by the 1960s appeared *Dialogical logic* developed by Paul Lorenzen and Kuno Lorenz, as a solution to some of the problems that arose in Lorenzen's *Operative Logik*.[6] Herewith, the epistemic turn initiated by

[1] Cf. Sundholm (1998, 2009).

[2] Cf. Rahman et al. (2012, pp. vii–ix).

[3] Lorenzen (1955).

[4] Prawitz (1979). For recent discussions related to the topic of harmony, see Read (2008, 2010).

[5] Cf. Schröder-Heister (2008).

[6] The main original papers are collected in Lorenzen and Lorenz (1978). For an historical overview of the transition from operative logic to dialogical logic see Lorenz (2001). For a presentation about the initial role of the framework as a foundation for intuitionistic logic, see Felscher (1994). Other papers have been collected more recently in Lorenz (2010a, b).

the proof-theoretical one was tackled with the notion of games that provided the dynamic features of the traditional dialectical reasoning. Inspired by Wittgenstein's *meaning as use* the basic idea of the dialogical approach to logic is that the meaning of the logical constants is given by the norms or rules for their use. On the other hand, a bit later on, still in the sixties, Jaakko Hintikka developed *game-theoretical semantic* (GTS). GTS is an approach to formal semantics that, like in the dialogical framework, grounds the concepts of truth or validity on game-theoretical concepts, such as the existence of a winning strategy for a player, though differently to the dialogical framework it is build up on the notion of model.[7] Furthermore, Hintikka combined the model-theoretical, the epistemic and the game-based traditions by means of the development of what is now known as *explicit epistemic logic*, where the epistemic content is introduced into the object language as an operator (of some specific modal kind) which yields propositions from propositions rather than as meaning conditions on the notion of proposition and inference. These kinds of operators were rapidly generalized covering several propositional attitudes including notably knowledge and belief.

These new impulses experienced, by 1980, a parallel renewal in the fields of theoretical computer science, computational linguistics, artificial intelligence and the formal semantics of programming languages. The impulse was triggered by the work of Johan van Benthem[8] and collaborators in Amsterdam who not only looked thoroughly at the interface between logic and games but also provided new and powerful tools to tackle the issue of the *expressivity* of a language – in particular the capability of propositional modal logic to express some decidable fragments of first-order logic.[9] New results in linear logic by J-**Y**. Girard in the interfaces between mathematical game theory and proof theory on one hand and argumentation theory and logic on the other hand resulted in the work of many others, including S. Abramsky, J. van Benthem, A. Blass, H. van Ditmarsch, D. Gabbay, M. Hyland, W. Hodges, R. Jagadessan, G. Japaridze, E. Krabbe, L. Ong, H. Prakken, G. Sandu D. Walton, and J. Woods who placed game semantics in the center of new concept of logic in which logic is understood as a dynamic instrument of inference.[10] A *dynamic turn*, as van Benthem puts it, is taking place and K. Lorenz's work on dialogical logic is a landmark in this turn. In fact, Lorenz's work can be more accurately described as the *dialogical turn* that re-established the link between dialectical reasoning and inference interaction.[11]

[7] Hintikka (1962, 1973, 1996a), Hintikka and Sandu (1997). See also Hintikka (1999) and in particular Hintikka et al. (1999). Shahid Rahman and Tero Tulenheimo (2009) studied the relation between dialogical logic and GTS. See also Tulenheimo (2011).

[8] van Benthem (1996, 2011, 2014).

[9] van Benthem (2001).

[10] See also: Blass (1992), Abramsky and Mellies (1999), Girard (1999), Lecomte and Quatrini (2010,2011), Lecomte (2011) and Lecomte and Tronçon (2011).

[11] This link provides the basis of a host of current and ongoing works in the history and philosophy of logic, going from the Indian, the Chinese, the Greek, the Arabic, the Hebraic traditions, the Obligationes of the Middle Ages to the most contemporary developments in the study of epistemic interaction. The main original papers on the dialogical approach are collected in Lorenzen and

Now, most of the logicians who endorse the dynamic turn[12] seem to ignore a recent study that represents a major advance in the task of recovering the logic of knowledge, namely, the development by Per Martin-Löf of the logical foundations of constructive mathematics that yielded Constructive Type Theory (CTT). This theory that provides a type-theoretical development of the Curry-Howard-Isomorphism between propositions as programs and propositions as sets-types, by introduction dependent types leads to the formulation of a fully-interpreted language – a language with content that challenges the usual metalogical approach to meaning of standard model-theoretical semantics.[13] In the CTT-framework the distinction between the classical (and realist) and constructivist positions can be expressed by distinguishing between assertion, the content of an assertion and the proposition expressed by an assertion in the following way: A proposition is classically determined by its truth-condition and constructively by its proof-objects; the content of an assertion amounts, according to the classical perspective, to the (transcendent) satisfaction of the truth-condition while constructively it amounts to the existence of a proof-object, and finally for both, classical and constructivists, an assertion indicates that the fact (=: truth-condition/proof-object) expressed by its content is known.

From the point of view of the modal approaches to epistemic logic in the Hintikka-style as developed for example by the school of J. Van Benthem at Amsterdam, the lack of interest in CTT is not a surprise, after all their game-theoretical approach is based on a model-theoretical semantics, where meaning is explained by metalinguistic means that relate uninterpreted signs and world. However, the, up to now, missing interface between the dialogical framework and CTT is particularly striking because of the common philosophical grounds of dialogical logic and those of constructive logic, where, as mentioned above, meaning of a linguistic expression is conceived as being constituted by the norms or rules for its use. Indeed, if the *use*-approach to meaning is intended to implement

Lorenz (1978). See also Kamlah and Lorenzen (1972, 1984) and Lorenzen and Schwemmer (1975). For an historical overview see Lorenz (2001). For a presentation about the initial role of the framework as a foundation for intuitionistic logic, see Felscher (1985). Other papers have been collected more recently in Lorenz (2008, 2010a, b). A detailed account of recent developments since, say, Rahman (1993), can be found in Rahman and Keiff (2005) and Keiff (2009). For the underlying metalogic see Clerbout (2014a, b). For a textbook presentation: Redmond and Fontaine (2011) and Rückert (2011a). For the key role of dialogic in regaining the link between dialectics and logic, see Rahman and Keff (2010). Keiff (2004a, b) and Rahman (2009) study Modal Dialogical Logic. Fiutek et al. (2010) study the dialogical approach to belief revision. Clerbout et al. (2011) studied Jain Logic in the dialogical framework. Popek (2012) develops a dialogical reconstruction of medieval *obligationes*. For other books see Redmond (2010) – on fiction and dialogic – Fontaine (2013) – on intentionality, fiction and dialogues – and Magnier (2013) – on dynamic epistemic logic van Ditmarsch et al. (2007) and legal reasoning in a dialogical framework.

[12]With the remarkable exceptions of of J.-Y. Girard and A. Ranta.

[13]Indeed, constructive type-theoretical grammar Ranta (1994), Ginzburg (2012) has now been successfully applied to the foundations of mathematics, logic, philosophy of logic, computer sciences, and to the semantics of natural languages. Particularly interesting is the fact that Ginzburg deploys CTT in order to capture the meaning of interaction underlying conversations in natural language.

in logic Wittgenstein's notion of language-games, who rejected the metalogical approach of model-theoretical semantics, the links between CTT and dialogical logic, seem to be very natural. More generally, if, once more, meaning is related to actions and those actions are understood as deploying games of answers and questions that involve the meaning of one main sentence, the game-theoretical approach to CTT follows naturally. One possible way to put it is to follow Mathieu Marion's[14] proposal and to make use of Robert Brandom's (1994, 2000) pragmatist take on inferentialism, which is led by two main insights of Kantian origin and one that stems from Brandom's reading of Hegel

1. That judgements are the fundamental units of knowledge, and
2. That human cognition and action is characterized by certain sorts of normative assessment.[15]
3. Communication is mainly conceived as cooperation in a joint social activity rather than on sharing contents.[16]

The crucial point of the epistemic approach, as mentioned above, is that assertion or judgement amounts to a knowledge claim and this is independent of classical or intuitionistic views cf. Prawitz (2012, p. 47). So, if meaning of an expression is deployed from its role in assertions, then an epistemic approach to meaning results. In relation to the second point, according to Brandom, the normative aspect is implemented via W. Sellar's notion of games of *giving and asking for reasons*, which deploy the intertwining of *commitments and entitlements*. Indeed, on Brandom's view, it is the chain of commitments and entitlements in a game of giving and asking for reasons that tights up judgement and inference.[17] Göran Sundholm (2013) provides the following formulation of the notion of inference in

[14]In fact, Mathieu Marion (2006, 2009, 2010) was the first to propose a link between Brandom's pragmatist inferentialism and dialogical logic in the context of Wilfried Hodges (2001, 2004(rev.2013)) challenges to the game-theoretical approaches. Moreover, another relevant antecedent of the present work is the PHD-thesis of Laurent Keiff (2007) who provided a thorough formulation of dialogical logic within the framework of speech-act theory.

[15]The normative aspect, rooted on the shift from Cartesian *certainty* to *bindingness of rules* distinguishes Brandom's pragmatism of others:

> *One of the strategies that guided this work is a commitment to the fruitfulness of shifting theoretical attention from the Cartesian concern with the grip we have on concepts – for Descartes, in the particular form of the centrality of the notion of* ***certainty*** [...] *– to the Kantian concern with the grip concepts have on us, that is the notion of necessity as the bindingness of the rules (including inferential ones) that determine how it is correct to apply those concepts* (Brandom 1994, p. 636).

[16]In relation to the model of holistic communication envisaged, Brandom (1994, p. 479) writes:

> *Holism about inferential significances has different theoretical consequences depending on whether one thinks of communication in terms of* ***sharing*** *a relation to one and the same* ***thing*** *(grasping a common meaning) or in terms of* ***cooperating*** *in a* ***joint activity*** [...].

[17]Moreover, according to Brandom, games of asking for reasons and giving them constitute the base of any linguistic practice:

a communicative context that can be also seen as describing the core of Brandom's pragmatist inferentialism:

> *When I say* "Therefore" ***I give others my authority for asserting the conclusion, given theirs for asserting the premsisses***.[18]

This is quite close to the main tenet of the dialogical approach to meaning with one important and crucial difference: though the pragmatist approach to meaning of the dialogical framework shares with Brandom's pragmatist inferentialism the claim that the meaning of linguistic expressions is related to their role in games of questions and answers and also endorses Brandom's notion of justification of a judgement as involving the interaction of commitments and entitlements, dialogicians maintain that more fundamental lower-levels should be distinguished. Those lower-level semantic levels include (i) the description of how to formulate a suitable question to give a posit and how to answer it, and (ii) the development of plays, constituted by several combinations of sequences of questions and answers brought forward as responses to the posit of a thesis. From the dialogical perspective, the level of judgements corresponds to the final stage of the chain of interactions just mentioned. More precisely, the justifications of judgements correspond to the level of winning strategies, that select those plays that turn out to be relevant for the drawing of inferences. Furthermore, as our discussion of the Axiom of Choice shows, the game-theoretical take on the dependent types is rooted on choices dependences, that can be seen as a result of the intertwining of games of questions and answers.

Let us point out that the distinctions: local meaning, play level and strategy level, drawn within the dialogical framework, seem to provide an answer to Brandom's question involving his claim that the "grasp of concepts" amounts to the mastery of inferential roles but this

> [. . .] *does not mean that in order to count as grasping a particular concept an individual must be disposed to make or otherwise endorse in practice all the right inferences involving it. To be in the game at all, one must make* ***enough*** *of the right moves — but how much is enough is quite flexible.* Brandom (1994, p. 636).

Indeed, from the dialogical point of view, in order to grasp the meaning of an expression, the individual must not need to know the moves that lead on how to win, he must not have a winning strategy, what it is required is that the knows what are the relevant moves he is entitled and committed to (local meaning) in order to

---

> *Sentences are expressions whose unembedded utterance performs a speech act such as making a claim, asking a question, or giving a command. Without expressions of this category; there can be no speech acts of any kind, and hence no specifically linguistic practice* (Brandom 2000, p. 125).

[18] Actually, Sundholm bases his formulation on J. L. Austin remarks in the celebrated paper of 1946, *Other Minds* rather than on Brandom's work. See also Sundholm (2009).

develop a play – in a similar way to knowing how to play chess does not necessarily mean to actually be in possession of a winning strategy. Knowing how to play allows to know what can count as a winning strategy, when there is one: strategic legitimacy (*Geltung*) is not to be found at the level of meaning-explanation. Thus, one way to see the motivations that animates the proposal to link CTT and games is to furnish the technical elements that bind the pragmatist approach to the grasp of concepts in Brandom's style, with the proof-theoretical CTT take on meaning.

The issue is now on how to link precisely the dynamic and epistemic turn with the fully-interpreted-approach of CTT in such a way that

(a) it incorporates the game-theoretical interpretation, where different kind of dependences are understood as deploying specific forms of interaction
(b) it makes it possible to express both the dynamics of knowledge acquisition and of meaning formation at the object language level.

It is our main claim that this can be achieved by the recent dialogical approach to CTT, where a language with content is developed that is able to meet the challenges of a framework where meaning and knowledge are conceived as constituted within interaction.[19] We will discuss two main cases, that represent two of the most cherished examples of Hintikka, namely the case of the Axiom of Choice and the study of anaphora (also one of Brandom's favourite subjects of study). To say it straight away, our claim is that the targets expressed by a) and b) can be achieved if we adopt the point of view that those functions that Hintikka identified as the ones that provide meaning to quantifier dependences are in fact object language proof-objects of the propositions in which the quantifiers occur, more precisely this functions are nothing more than *dependent proof-objects*. Moreover, proof-objects are made of more elementary constituents that we call *play-objects*. However all this considerations seem to point out, that at the end Hintikka's claim of *super-classicality* is not compatible with a theory of meaning that makes interaction explicit at the object language level.

3.2 The Dialogical Approach to CTT[20]

3.2.1 Dialogical Logic and the Pragmatist Theory of Meaning

The dialogical approach to logic is not a specific logical system but rather a rule-based semantic framework in which different logics can be developed, combined and compared. An important point is that the rules that fix meaning can be of more

[19]Cf. Rahman and Clerbout (2013, 2015), Clerbout and Rahman (2015), Jovanovic (2013).

[20]The present overview on the dialogical approach to CTT is based on Clerbout and Rahman (2015) – see also Rahman and Clerbout (2013, 2015), Rahman et al. (2014).

than one kind. This feature of its underlying semantics quite often motivated the dialogical framework to be understood as a *pragmatist* semantics. More precisely, in a dialogue two parties argue about a thesis respecting certain fixed rules. The player that states the thesis is called Proponent (**P**), his rival, who contests the thesis is called Opponent (**O**). In its original form, dialogues were designed in such a way that each of the plays end after a finite number of moves with one player winning, while the other loses. Actions or moves in a dialogue are often understood as speech-acts involving *declarative utterances or posits and interrogative utterances or requests*. The point is that the rules of the dialogue do not operate on expressions or sentences isolated from the act of uttering them. The rules are divided into particle rules or rules for logical constants (*Partikelregeln*) and structural rules (*Rahmenregeln*). The structural rules determine the general course of a dialogue game, whereas the particle rules regulate those moves (or utterances) that are requests and those moves that are answers (to the requests).[21]

Crucial for the dialogical approach are the following points[22]:

1. The distinction between *local* (rules for logical constants) and *global* meaning (included in the structural rules that determine how to play)
2. The player independence of local meaning
3. The distinction between the play level (local winning or winning of a play) and the strategic level (existence of a winning strategy).
4. A notion of validity that amounts to winning strategy *independently of any model* instead of winning strategy for *every* model.
5. The distinction between non formal and formal plays – neither the latter nor the first kind concerns plays where the actions of positing an elementary sentences require a meta-language level that provides their truth.

Recent developments in dialogical logic show that the CTT approach to meaning is very natural to game-theoretical approaches where (standard) metalogical features are explicitly displayed at the object language-level.[23] Thus, in some way, this vindicates, albeit in quite of a different manner, Hintikka's plea for the fruitfulness of game-theoretical semantics in the context of epistemic approaches to logic, semantics and the foundations of mathematics. In fact, from the dialogical point of view, those actions that constitute the meaning of logical constants, such as choices, are a crucial element of its full-fledged (local) semantics. Indeed, if meaning is conceived as being constituted during interaction, then all of the actions involved in the constitution of the meaning of an expression should be rendered explicit. They should all be part of the object language. The roots of this perspective are based on Wittgenstein's *Unhintergehbarkeit der Sprache* – one of the tenets of Wittgenstein

[21]For a brief presentation of standard dialogical logic see appendix.

[22]Cf. Rahman (2012).

[23]Cf. Rahman and Clerbout (2013, 2015), Clerbout and Rahman (2015).

that Hintikka explicitly rejects.[24] According to this perspective of Wittgenstein language-games are purported to accomplish the task of displaying this "internalist feature of meaning". Furthermore, one of the main insights of Kuno Lorenz' interpretation of the relation between the so-called *first* and *second* Wittgenstein is based on a thorough criticism of the metalogical approach to meaning (Lorenz 1970, pp. 74–79).[25]

If we recall Hintikka's (1996b) extension of van Heijenoort's distinction of a *language as the universal medium* and *language as a calculus,* the point is that the dialogical approach shares some tenets of both conceptions. Indeed, on one hand the dialogical approach shares with universalists the view that we cannot place ourselves outside our language, on the other it shares with the anti-universalists the view that we can develop a methodical of *local* truth.

Similar criticism to the metalogical approach to meaning has been raised by G. Sundholm (1997, 2001) who points out that the standard model-theoretical semantic turns semantics into a meta-mathematical formal object where syntax is linked to meaning by the assignation of truth values to uninterpreted strings of signs (formulae). Language does not any more *express content* but it is rather conceived as a system of signs that speaks *about* the world – provided a suitable metalogical link between signs and world has been fixed. Moreover, Sundholm (2016, forthcoming) shows that the cases of quantifier-dependences that motivate Hintikka's IF-logic can be rendered in the frame of CTT. What we add to Sundholm's remark is that even the game-theoretical interpretation of these dependences can be given a CTT formulation, provided this is developed within a dialogical framework.

In fact, in his 1988 paper, Ranta linked for the first time game-theoretical approaches with CTT. Ranta took Hintikka's Game-Theoretical Semantics as a case

[24]Hintikka shares this rejection with all those who endorse model-theoretical approaches to meaning.

[25]In this context Lorenz writes:

> *Also propositions of the metalanguage require the understanding of propositions,* [...] *and thus can not in a sensible way have this same understanding as their proper object. The thesis that a property of a propositional sentence must always be internal, therefore amounts to articulating the insight that in propositions about a propositional sentence this same propositional sentence does not express anymore a meaningful proposition, since in this case it is not the propositional sentence that it is asserted but something about it.*
>
> *Thus, if the original assertion (i.e., the proposition of the ground-level) should not be abrogated, then this same proposition should not be the object of a metaproposition,* [...]. (Lorenz 1970, p.75).
>
> *While originally the semantics developed by the picture theory of language aimed at determining unambiguously the rules of "logical syntax" (i.e. the logical form of linguistic expressions) and thus to justify them* [...] *– now language use itself, without the mediation of theoretic constructions, merely via "language games", should be sufficient to introduce the talk about "meanings" in such a way that they supplement the syntactic rules for the use of ordinary language expressions (superficial grammar) with semantic rules that capture the understanding of these expressions (deep grammar).* (Lorenz 1970, p.109).

study, though his point does not depend on that particular framework: in game-based approaches, a proposition is a set of winning strategies for the player positing the proposition.[26] In game-based approaches, the notion of truth is at the level of such winning strategies. Ranta's idea should therefore let us safely and directly apply to instances of game-based approaches methods taken from constructive type theory.

But from the perspective of game-theoretical approaches, reducing a game to a set of winning strategies is quite unsatisfactory, especially when it comes to a theory of meaning. This is particularly clear in the dialogical approach in which different levels of meaning are carefully distinguished. There is thus the level of strategies which is a level of meaning analysis, but there is also a level prior to it which is usually called the level of plays. The role of the latter level for developing an analysis is crucial according to the dialogical approach, as pointed out by Kuno Lorenz in his 2001 paper:

> [...] *for an entity* [A] *to be a proposition there must exist a dialogue game associated with this entity* [...] *such that an individual play of the game where A occupies the initial position* [...] *reaches a final position with either win or loss after a finite number of moves* [...]

For this reason we would rather have propositions interpreted as sets of what we shall call play-objects and read the expression

$$p : \varphi$$

as "p is a play-object for φ".

Thus, Ranta's work on proof-objects and strategies is the end, not the beginning, of the dialogical project.

3.2.2 The Formation of Propositions

Before delving into the details about play-objects, let us first discuss the issue of forming expressions and especially propositions in the dialogical approach.

It is presupposed in standard dialogical systems that the players use well-formed formulas (wff). The well formation can be checked at will, but only with the usual meta reasoning by which the formula is checked to indeed observe the definition of a wff. We want to enrich the system by first allowing players to enquire on the status of expressions and in particular to ask if a certain expression is a proposition. We thus start with dialogical rules explaining the formation of propositions. These rules are local rules which are added to the particle rules giving the local meaning of logical constants (see next section).

[26] That player can be called Player 1, Myself or Proponent.

A remark before displaying the formation rules: because the dialogical theory of meaning is based on argumentative interaction, dialogues feature expressions which are not only posits of sentences. They also feature requests, used for challenges, as the formation rules below and the particle rules in the next section illustrate. Because of the *no entity without type* principle, it seems at first glance that we should specify the type of these actions during a dialogue: the type "*formation-request*". It turns out we should not: an expression such as "$?_F$*: formation-request*" is a judgement that some action $?_F$ is a formation-request, which should not be confused with the actual act of requesting. We also consider that the force symbol $?_F$ makes the type explicit. Hence the way requests are written in rules and dialogues in this work.

The formation rules are given in the following table. Notice that a posit '$\bot$: prop' cannot be challenged: this is the dialogical account of the fact that the falsum $\bot$ is by definition a proposition.

Posit	Challenge [when different challenges are possible, the challenger chooses]	Defence
X *!* Γ *: set*	**Y** $?_{can}\Gamma$	**X** *!* a_1*:* Γ, **X** *!* a_2*:* Γ, ...
	or	(**X** provides the canonical elements of Γ)
	Y $?_{gen}\Gamma$	**X** *!* a_i*:* $\Gamma \Rightarrow a_j$*:* Γ
	or	(**X** provides a generation method for Γ)
	Y $?_{eq}\Gamma$	(**X** provides the equality rule for Γ[a])
X *!* $\varphi \vee \psi$ *: prop*	**Y** $?_{F\vee 1}$	**X** *!* φ *: prop*
	or	
	Y $?_{F\vee 2}$	**X** *!* ψ *: prop*
X *!* $\varphi \wedge \psi$ *: prop*	**Y** $?_{F\wedge 1}$	**X** *!* φ *: prop*
	or	
	Y $?_{F\wedge 2}$	**X** *!* ψ *: prop*
X *!* $\varphi \rightarrow \psi$ *: prop*	**Y** $?_{F\rightarrow 1}$	**X** *!* φ *: prop*
	or	
	Y $?_{F\rightarrow 2}$	**X** *!* ψ *: prop*
X *!* $(\forall x : A)\ \varphi(x)$ *: prop*	**Y** $?_{F\forall 1}$	**X** *!* A *: set*
	or	
	Y $?_{F\forall 2}$	**X** *!* $\varphi(x)$ *: prop* $(x : A)$
X *!* $(\exists x : A)\ \varphi(x)$ *: prop*	**Y** $?_{F\exists 1}$	**X** *!* A *: set*
	or	
	Y $?_{F\exists 2}$	**X** *!* $\varphi(x)$ *: prop* $(x : A)$
X *!* $B(k)$ *: prop* (for atomic B)	**Y** $?_F$	**X** *sic* (n) (**X** indicates that **Y** posited it at move n)
X *!* $\bot$ *: prop*	–	–

[a]Equality rules are presented in the next section

The next rule is not a formation rule per se but rather a substitution rule.[27] When φ is an elementary sentence, the substitution rule helps explaining the formation of such sentences.

3.2.2.1 Posit-Substitution

When a list of variables occurs in a posit with proviso, the challenger **Y** can ask **X** to substitute those variables: he does so by positing an instantiation of the proviso, in which he (**Y**) is the one who chooses the instantiations for the variables.[28]

Posit	Challenge	Defence
X! $\pi(x_1, \ldots, x_n)\ (x_i : A_i)$	**Y**! $a_1 : A_1, \ldots, a_n : A_n$	**X**! $\pi(a_1 \ldots a_n)$

A particular case of posit substitution is when the challenger simply posits the whole assumption as it is without introducing new instantiation terms. This is particularly useful in the case of formation plays: see an application in move 5 of the second example below.

Posit	Challenge	Defence
X! $\pi(x_1, \ldots, x_n)\ (x_i : A_i)$	**Y**! $x_1 : A_1, \ldots, x_n : A_n$	**X**! $\pi(x_1, \ldots, x_n)$

Remarks on the Formation Dialogues

(a) Conditional formation posits:
A crucial feature of formation rules is that they enable the displaying of the syntactic and semantic presuppositions of a given thesis which can thus be examined by the Opponent before running the actual dialogue on the thesis. For instance if the thesis amounts to positing φ, then before launching an attack, the Opponent can ask for its formation. Defending on the formation of φ might bring the Proponent to posit that φ is a proposition, provided that A, for instance, is a set is conceded. In this situation the Opponent might concede A is a set, but only after the Proponent displayed the constitution of A.

(b) The formation of elementary sentences and the response *sic*(n)
It might look as if the affirmation *sic*(n) as responding to the formation request for a given elementary expression is unsatisfactory. However, in fact it expresses

[27] It is an application of the original rule from CTT given in Ranta (1994, p.30).

[28] More precisely, in the case where the defender did not commit himself to the proviso, the dialogical approach allows a distinction here discussed in the next section.

the fact, that, if such a move is possible, that well-formation has been already been examined before and consequently conceded by the opponent. Take the case that the affirmation is $A(b)$: prop, in such a case, either the Proponent must show how this is obtained of the underlying propositional function, or he can recall that this has been already examined before. This is what the response *sic(n)* expresses.

By way of illustration, here is an example where the Proponent posits the thesis $(\forall x : A)(B(x) \rightarrow C(x))$: prop given that A : set, $B(x)$: prop $(x : A)$ and $C(x)$: prop $(x : A)$, where the three provisos appear as initial concessions by the Opponent.[29] Normally we should give all the rules of the game before giving an example, but we make an exception here because the standard structural rules of appendix are enough to understand the following plays. We can focus this way on illustrating the way formation rules can be used.

	O			P	
I	*! A : set*				
II	*! B(x) : prop (x : A)*				
III	*! C(x) : prop (x : A)*				
				! (∀x : A) B(x) → C(x) : prop	0
1	*n := 1*			*m := 2*	2
3	$?_{F\forall 1}$	(0)		*! A : set*	4

Explanations

- I to III: ***O*** concedes that A is a set and that $B(x)$ and $C(x)$ are propositions provided x is an element of A,
- Move 0: ***P*** posits that the main sentence, universally quantified, is a proposition (under the concessions made by ***O***),
- Moves 1 and 2: the players choose their repetition ranks,[30]
- Move 3: ***O*** challenges the thesis by asking the left-hand part as specified by the formation rule for universal quantification,
- Move 4: ***P*** responds by positing that A is a set. This has already been granted with the premise I so even if ***O*** were to challenge this posit, the Proponent could refer to this initial concession. Later on, we will introduce the structural rule SR3 to deal with this phenomenon. Thus ***O*** has no further possible move, the dialogue ends here and is won by ***P***.

[29]The example stems from Ranta (1994, p.31).

[30]The device of repetition rank is introduced in the structural rules which we present in the appendix. See also Clerbout (2014a, b, c) for detailed explanations on this notion.

Obviously, this dialogue does not cover all the aspects related to the formation of $(\forall x : A)\ B(x) \rightarrow C(x)$: prop. Notice however that the formation rules allow an alternative move for the Opponent's move 3.[31] Hence another possible course of action for ***P*** arises.

	O			P	
I	*! A : set*				
II	*! B(x) : prop (x : A)*				
III	*! C(x) : prop (x : A)*				
				! (∀x : A) B(x) → C(x) : prop	0
1	*n := 1*			*m := 2*	2
3	$?_{F\forall 2}$	(0)		*! B(x) → C(x) : prop (x : A)*	4
5	*! x : A*	(4)		*! B(x) → C(x) : prop*	6
7	$?_{F\rightarrow 1}$	(6)		*! B(x) : prop*	10
9	*! B(x) : prop*		(II)	*! x : A*	8

Explanations

The second dialogue starts like the first one until move 2. Then:

- Move 3: This time **O** challenges the thesis by asking for the right-hand part,
- Move 4: **P** responds, positing that $B(x) \rightarrow C(x)$ is a proposition provided $x : A$,
- Move 5: **O** uses the substitution rule to challenge move 4 by granting the proviso,
- Move 6: **P** responds by positing that $B(x) \rightarrow C(x)$ is a proposition,
- Move 7: **O** then challenges move 6 by asking the left-hand part as specified by the formation rule for material implication.

To defend this **P** needs to make an elementary move. But since **O** has not played it yet, **P** cannot defend it at this point. Thus:

- Move 8: **P** launches a counterattack against assumption II by applying the substitution rule,
- Move 9: **O** answers to move 8 and posits that $B(x)$ is a proposition,
- Move 10: **P** can now defend in reaction to move 7 and win this dialogue.

Then again, there is another possible path for the Opponent because she has another possible choice for her move 7, namely asking the right-hand part. This yields a dialogue similar to the one above except that the last moves are about $C(x)$ instead of $B(x)$.

[31] As a matter of fact increasing her repetition rank would allow her to play the two alternatives for move 3 within a single play. But increasing the Opponent's rank usually yields redundancies (Clerbout 2014a, b) making things harder to understand for readers not familiar with the dialogical approach. Hence our choice to divide the example into different simple plays.

By displaying these various possibilities for the Opponent, we have entered the *strategical* level. This is the level at which the question of the good formation of the thesis gets a definitive answer, depending on whether the Proponent can always win – i.e., whether he has a winning strategy. We have introduced the basic notions related to this level in the previous section. See also the appendix and Clerbout and Rahman (2015, Chaps. 3 and 5).

Now that the dialogical account of formation rules has been clarified, we may further develop our analysis of plays by introducing play-objects.

3.2.3 Play-Objects

The idea now is to design dialogical games in which the players' posits are of the form "$p : \varphi$" and give their meaning by the way they are used in the game: how they are challenged and defended. This requires analysing the form of a given play-object p, which depends on φ, and how a play-object can be obtained from other, simpler, play-objects. The standard dialogical semantics appendix for logical constants gives us the information we need. The main logical constant of the expression at stake provides the basic information as to what a play-object for that expression consists of:

A play for $\boldsymbol{X}\ !\varphi \vee \psi$ is obtained from two plays p_1 and p_2, where p_1 is a play for $\boldsymbol{X}\ !\varphi$ and p_2 is a play for $\boldsymbol{X}\ !\psi$. According to the standard dialogical approach to disjunction, the player $\boldsymbol{X}$ is the one who can switch from p_1 to p_2 and conversely.

A play for $\boldsymbol{X}\ !\varphi \wedge \psi$ is obtained similarly, except that the player $\boldsymbol{Y}$ is the one who can switch from p_1 to p_2.

A play for $\boldsymbol{X}\ !\varphi \rightarrow \psi$ is obtained from two plays p_1 and p_2, where p_1 is a play for $\boldsymbol{Y}\ !\varphi$ and p_2 is a play for $\boldsymbol{X}\ !\psi$. The player $\boldsymbol{X}$ is the one who can switch from p_1 to p_2.

The standard dialogical particle rule for negation interprets $\neg\varphi$ as an abbreviation for $\varphi \rightarrow \bot$, although it is usually left implicit. From this follows that one obtains plays for $\boldsymbol{X}\ !\neg\varphi$ in a similar way to plays for material implication, that is from two plays p_1 and p_2, where p_1 is a play for $\boldsymbol{Y}\ !\varphi$, p_2 is a play for $\boldsymbol{X}\ !\bot$, and $\boldsymbol{X}$ can switch from p_1 to p_2. Notice that this approach covers the standard game-theoretical interpretation of negation as role-switch: p_1 is a play for a $\boldsymbol{Y}$-move.

As for quantifiers, a detailed discussion will be given after the particle rules. We would like to point out for now that, just like what is done in constructive type theory, we are dealing with quantifiers for which the type of the bound variable is always specified. We thus consider expressions of the form (Qx: A)φ, where Q is a quantifier symbol.

The table on next page presents the particle rules.

Posit	Challenge	Defence
X! φ	**Y**? *play-object*	X ! $p : \varphi$
(where no play-object has been specified for φ)		
X! $p : \varphi \vee \psi$	**Y**?$_{prop}$	**X**!$\varphi \vee \psi$: *prop*
		X! $L^{\vee}(p) : \varphi$
	Y?[φ/ψ]	Or
		X! $R^{\vee}(p) : \psi$
		[the defender has the choice]
X! $p : \varphi \wedge \psi$	**Y**?$_{prop}$	**X**! $\varphi \wedge \psi$: *prop*
	Y?$_L$	**X**! $L^{\wedge}(p) : \varphi$
	Or	respectively
	Y?$_R$	**X**! $R^{\wedge}(p) : \psi$
	[the challenger has the choice]	
X! $p : \varphi \rightarrow \psi$	**Y** ?$_{\text{prop}}$	**X**! $\varphi \rightarrow \psi$: *prop*
	Y! $L^{\rightarrow}(p) : \varphi$	**X**! $R^{\rightarrow}(p) : \psi$
X! $p : \neg\varphi$	**Y**?$_{prop}$	**X**! $\neg\varphi$: *prop*
	Y! $L^{\perp}(p) : \varphi$	**X**! $R^{\perp}(p) : \perp$
X! $p : (\exists x : A)\varphi$	**Y**?$_{prop}$	**X**! $(\exists x : A)\varphi$: *prop*
	Y?$_L$	**X**! $L^{\exists}(p) : A$
	Or	Respectively
	Y?$_R$	**X**! $R^{\exists}(p) : \varphi(L(p))$
	[the challenger has the choice]	
X! $p : \{x : A \mid \varphi\}$	**Y**?$_L$	**X**! $L^{\{\cdots\}}(p) : A$
	Or	Respectively
	Y?$_R$	**X**! $R^{\{\cdots\}}(p) : \varphi(L(p))$
	[the challenger has the choice]	
X! $p : (\forall x : A)\varphi$	**Y**?$_{prop}$	**X**! $(\forall x : A)\varphi$: *prop*
	Y! $L^{\forall}(p) : A$	**X**! $R^{\forall}(p) : \varphi(L(p))$
X! $p : B(k)$ (for atomic B)	**Y**?$_{prop}$	**X**! $B(k)$: *prop*
	Y?	**X***sic (n)*
		(**X** indicates that **Y** posited it at move n)

Let us point out that we have added a challenge of the form ***Y*** ?$_{\text{prop}}$ by which the challenger questions the fact that the expression at the right-hand side of the semicolon is a proposition. This connects back with the formation rules of the preceding section via ***X***'s defence. Further details will be given in the discussion after the structural rules.

It may happen that the form of a play-object is not explicit at first. In such cases we deal with expressions of the form, e.g., "$p : \varphi \wedge \psi$". In the relevant challenges and defences, we then use expressions such as $L^{\wedge}(p)$ and $R^{\wedge}(p)$ used in our example. We call these expressions *instructions*. Their respective interpretations are "take the left part of p" and "take the right part of p". In instructions we indicate the logical constant at stake[32]: it keeps the formulations explicit enough, in particular in the case of embedded instructions. We must also keep in mind the important differences between play-objects depending on the logical constant that is used. Consider for example the case of conjunction and disjunction:

- A play-object p for a disjunction is composed by two play-objects, but each of them constitutes a sufficient play-object for the disjunction. Moreover it is the defender who makes the choice between $L^{\vee}(p)$ and $R^{\vee}(p)$.
- A play-object p for a conjunction is also composed by two play-objects, but this time the two of them are necessary to constitute the one for the conjunction. It is then the challenger's privilege to ask for either or both (provided the other rules allow him to do so).

Accordingly, $L^{\wedge}(p)$ and $L^{\vee}(p)$, say, are actually different things and the notation takes that into account.

Let us now focus on the quantifier rules. There are two distinct moments in the meaning of quantifiers, brought out by dialogical semantics: choosing a suitable substitution term for the bound variable, and instantiating the formula after replacing the bound variable with the chosen substitution term. However the standard dialogical approach tends to presuppose a unique and global collection of objects on which the quantifiers range. Things are different with the explicit language borrowed from CTT. Quantification is always relative to a set, and there are sets of many different kinds of objects (for example: sets of individuals, sets of pairs, sets of functions, etc). Owing to the instructions we can give a general form for the particle rules, and the object is specified in a third and later moment, when instructions are "resolved" by means of the structural rule SR4.1 displayed in the next section.

Constructive type theory clearly shows the basic similarity there is between conjunction and existential quantifier on the one hand and material implication and universal quantifier on the other hand, as soon as propositions are thought of as sets. Briefly, the point is that they are formed in similar ways and their elements are generated by the same kind of operations.[33] In our approach, this

[32] If needed, we use subscripts to prevent scope ambiguities in the case of embedded occurrences of the same quantifier.

[33] More precisely, conjunction and existential quantifier are two particular cases of the Σ operator (disjoint union of sets), whereas material implication and universal quantifier are two particular cases of the Π operator (indexed product on sets). See for example Ranta (1994, chapter 2).

similarity manifests itself in the fact that a play-object for an existentially quantified expression is of the same form as a play-object for a conjunction. Similarly, a play-object for a universally quantified expression is of the same form as one for a material implication.[34]

The particle rule just before the one for universal quantification is a novelty in the dialogical approach. It involves expressions commonly used in Constructive Type Theory to deal with separated subsets. The idea is to understand those elements of A *such that* φ as expressing that at least one element $L^{\{\cdots\}}(p)$ of A witnesses $\varphi(L^{\{\cdots\}}(p))$. The same correspondence that linked conjunction and existential quantification now appears.[35] This is not surprising since such posits actually have an existential aspect: in $\{x : A \mid \varphi\}$ the left part "$x : A$" signals the existence of a play-object. Let us point out that since the expression stands for a set, when $\boldsymbol{X}$ posits it, it is not presupposed to be a proposition. This is why it cannot be challenged with the request "$?_{\text{prop}}$".

As we previously said, in the dialogical approach to CTT every object is known as instantiating a type and this constitutes the most elementary form of assertion $a : A$. Furthermore, instructions are in fact substitution commitments close to the sense mentioned in the above quote. A thorough study is yet to be done on the substitutional approach to subsentential expressions and the role of instructions, though it would be necessary in our view for the exploration of both the formal consequences of Brandom's insights and the philosophical tenets underlying the notion of instruction.

Let us now consider the rule for the elementary case. In this rule, but also in the associated formation rule of the previous section, the defence "$sic(n)$" recalls that the adversary has previously made the same posit. The rule works in a similar fashion as the formal rule of the standard formulation (see appendix), except that it is applicable to both players: it is not limited to the Proponent. We say similar in the sense that the rule allows players to perform a kind of copy-cat. Once that aspect of the formal rule is settled, we can work with a modified version of the rule which we will introduce with more explanations in the next section.

[34] Still, if we are playing with classical structural rules, there is a slight difference between material implication and universal quantification which we take from Ranta (1994, Table 2.3), namely that in the second case p_2 always depends on p_1.

[35] As pointed out in Martin-Löf (1984), subset separation is another case of the Σ operator. See in particular p.53:

> *Let A be a set and B(x) a proposition for x∈A. We want to define the set of all a∈A such that B(a) holds (which is usually written {x∈A : B(x)}). To have an element a∈A such that B(a) holds means to have an element a∈A together with a proof of B(a), namely an element b∈B(a). So the elements of the set of all elements of A satisfying B(x) are pairs (a,b) with b∈B(a), i.e., elements of* (Σx∈A)B(x). *Then the Σ-rules play the role of the comprehension axiom (or the separation principle in ZF).*

Despite the similarity we have just mentioned, there is a crucial difference with standard dialogical games. Elementary sentences are associated with play-objects, and one such sentence can be associated with many different play-objects in actual courses of the game. Therefore, and this is a most important point, the defence "*sic*(*n*)" does not express a copy-cat on the elementary sentence alone, but on the whole posit. We thus have a game rule such that, for a given elementary sentence, there are as many ways to give reasons for it (to defend it) as there are play-objects for it. Formulating the rule with the defence "*sic*(*n*)" is very different from merely integrating the standard formal rule at the local level: "*sic*(*n*)" is an abbreviation useful to provide an abstract rule, but because play-objects are introduced, it actually embodies a fully fledged semantics in terms of asking for and giving reasons. See Clerbout and Rahman (2013).

So far, apart from the rule for subset-separation and the rule for elementary sentences, we have mostly adapted the rules of standard dialogical games to the explicit language we are working with. Now because of the explicit nature of this language, there are more rules related to the meaning explanations of play-objects and types. The next rules involve what is known in CTT as *definitional equality*. These rules introduce a different kind of provisional clause, namely a clause in which the defender is the player committed to the expression within the clause and thus he, rather than the challenger, will eventually posit it. In standard CTT there is no need for such a distinction since there are no players. However, in dialogical games the distinction can and must be made depending on who posits the proviso. Accordingly we use of the notation $<\ldots>$ to signal that it is the player making the posit who is committed to the expression in the proviso clause and $(\ldots)$ when it is the opponent.

We have already considered the latter case in this section. Let π be a posit and $<\ldots>$ a proviso which the utterer is committed to. The general form of the rule for provisos is the following:

Posit	Challenge	Defence
$\mathbf{X}!\ \pi < \ldots >$	$\mathbf{Y}?_{[\pi]}$	$\mathbf{X}!\ [\pi]$
	Or	
	$\mathbf{Y}?_{[<\ldots>]}$	$\mathbf{X}![<\ldots>]$
	where $?[\pi]$ and $![\pi]$ stand respectively for the relevant challenge or defence against π, and similarly for $?[<>]$, $![<>]$	

In the initial posit, **X** commits himself to both π and the proviso. Hence **Y** is entitled to question either one, and he is the one to choose which to ask for. The rule states that the challenger can question either part of the initial posit, and that in each case he does so depending on the form of the expression. An illustration is helpful here. Assume the initial posit is $p : (\forall x : A)B(x)<c : C>$ which reads "given $c : C$ we have $B(x)$ for all $x : A$; and the player making the posit commits himself to the proviso". Then the rule is applied as described in the next table.

Posit	Challenge	Defence
X! $p :(\forall x : A)\, Bx <c : C>$	**Y**? $L^{\forall}(p) : A$	**X**! $R^{\forall}(p) : B(L(p))$
	Or	
	Y$?_{[c\,:\,C]}$	**X***sic (n)*
	where $?[\pi]$ and $![\pi]$ stand respectively for the relevant challenge or defence against π, and similarly for $?[<>], ![<>]$	

In this case, π involves universal quantification and the proviso is the elementary posit $c : C$. Thus, the first possible challenge for ***Y*** consists in applying the particle rule for universal quantification, whereas the second possible challenge is done by applying the rule for elementary posits. The possible defences by ***X*** are then in turn determined by these rules.

A typical case in which provisos of the form <...> occur is functional substitution. Assume some function f has been introduced, for example with $f(x) : B\ (x : A)$. When a player uses $f(a)$ in a posit, for some $a : A$, the antagonist is entitled to ask him what the output substitution-term of f is, given the substitution-term a as input. Now $f(a)$ can be used either at the left or at the right of the colon. Accordingly we have two rules:

(Function-substitution)

Posit	Challenge	Defence
X! $f(a) : \varphi$	**Y**$f(a)?_{<=>}$	**X**! $f(a)/ki : \varphi < f(a) = ki : B >$
X!$\alpha : \varphi[f(a)]$	**Y**$f(a)/?_{<=>}$	**X**! $\alpha : \varphi[f(a)/ki]$ $<\varphi[f(a)] = \varphi[f(a)/ki] : set >$

The subscript '<=>' in the challenges indicates that the substitution is related to some equality, and the defender endorses an equality in the proviso of the defence. The second rule - where α can be a play-object or an instruction – is applied in the dialogical take on the Axiom of Choice. See the "second play" in Sect. 3.3.

Important Remark These two rules express a double commitment for the defender who is committed to the proviso in the defence. One might therefore argue that the rules could also be formulated as involving two challenges (and two defences). There are however two problems with such an approach. For illustration purposes, let us consider such a formulation of the second rule involving two steps:

Posit	Challenge	Defence
X! $\alpha : \varphi[f(a)]$	**Y**! $L(f(a))/?$	**X**! $p : \varphi[f(a)/ki]$
	Y! $R(f(a))/?$	**X**! $\varphi[f(a)] = \varphi\ [f(a)/ki] : set$

The first problem is that the second challenge works as if the proviso $\varphi[f(a)]=\varphi[f(a)/k_i]$: set was implicit in the initial posit and had to be made explicit. However this is a slightly misguided approach since the proviso does not concern the initial posit: the proviso must be established only after $\boldsymbol{X}$ has chosen k_i for the substitution. The second problem is related to the first: in such a formulation the challenger is the one who can choose between asking $\boldsymbol{X}$ to perform the substitution and asking him to posit the proviso. It thus allows the challenger to perform just the second challenge without asking for the substitution, which brings us back to the first problem. Moreover, introducing a choice for one of the players results, when the rule can be applied, in multiplying the number of alternative plays (in particular when the repetition rank of the challenger is 1). For all these reasons, such an alternative formulation is less satisfactory than the one we gave above.

Functional substitution is closely related to the Π-Equality rule, which we now introduce together with $\Sigma-$Equality.

(Π-Equality) We use the CTT notation Π which covers the cases of universal quantification and material implication.

Posit	Challenge	Defence
$\mathbf{X}!\ p : (\Pi x : A)\varphi$		
$\mathbf{Y}!\ L^{\Pi}(p)/a : A$		
$\mathbf{X}!\ R^{\Pi}(p) : \varphi(a/x)$	$\mathbf{Y}?_{\Pi\text{-}Eq}$	$\mathbf{X}!\ p(a) = R^{\Pi}(p) : \varphi(a/x)$

(Σ-Equality) The rule is similar for existential quantification, subset separation, and conjunction. Thus we use the notation from CTT which uses the Σ operator. In the following rule I^{Σ} can be either L^{Σ} or R^{Σ}, and i can be either 1 or 2. Moreover, it is 1 when I is L and 2 when I is R.

Posit	Challenge	Defence
$\mathrm{X}\ !\ p : (\Sigma\ x : \varphi_1)\varphi_2$		
$\mathrm{Y}\ I^{\Sigma}(p)/?$		
$\mathrm{X}\ !\ p_i/I^{\Sigma}(p) : \varphi_i$	$\mathbf{Y}?_{\Sigma\text{-}Eq}$	$\mathbf{X}!\ I^{\Sigma}(p) = p_i\colon \varphi_i$

Notice that these rules have several preconditions: there is no lone initial posit triggering the application of the rule. From a dialogical perspective, these rules intend to allow the challenger to take advantage from information from the history of the current play – including resolutions of instructions – to make $\mathbf{X}$ posit some equality. For an application, see the second play in Sect. 3.3 where the Π-Equality rules play a prominent role.

These rules strongly suggest a close connection between the CTT equality rules for logical constants and the dialogical instructions through what we will call in the next section their *resolution*.

In fact, equality rules can be seen as making explicit the use of the formal rule in relation to the task of carrying out instructions. Hence, under this perspective, identities express explicitly some specific forms of interaction. Let us briefly discuss this point:

Assume that the Proponent brings forward the thesis that if the Opponent concedes the conjunction, say $A \wedge B$, he (the Proponent) will be able to successfully defend the assertion $B \wedge A$, that is, that **P** has a winning strategy for the commutative transformation of the conjunction. Let us present informally the dialogical development of this thesis:

1. **O** $! \; p : A \wedge B$ (concession)
2. **P** $! \; q : B \wedge A$
3. **O** $?_{\mathrm{L}}$ (the Opponent launches his challenge asking for the left component)
4. **P** $! \; L^{\wedge}(q) : B$
5. **O** $L^{\wedge}(q)/?$ (**O** asks **P** to carry out the instruction by picking out one play-object)
6. Since we are focusing on a winning strategy, we will assume that **P** makes the smartest move, and this is certainly to launch a counter-attack: the idea is to force **O** to choose a play-object first and then copy-cat it, before he goes on to answer the challenge of move 5:
   P $?_{\mathrm{R}}$
7. **O** $! \; R^{\wedge}(p) : B$
8. **P** $R^{\wedge}(p)/?$ (**P** asks **O** to carry out the instruction by picking out one play-object for the right side of the conjunction)
9. **O** $! \; b : B$ (**O** carries out the instruction by choosing the play-object b)
10. Now the Proponent has the information he needed, and copies the Opponents choice to answer **O**'s challenge stated at move 5:
    P $! \; b : B$
    (It should be clear that a similar end will happen if **O** starts by challenging the right component of the conjunction-posit)

Now, let us try to make explicit what happened. The point is that the Proponent is in fact considering the right part of p as definitionally equal to the left part of q. If we were to make explicit this move, the following definitional equality will come out:

$$R^{\wedge}(p) = L^{\wedge}(q) : B$$

The influence of the definitional equality of play-objects on the equality of propositions is exemplified at its best in the case of quantifiers. Take, for instance, the thesis that there is a **P**-winning strategy for $p : (\exists x : A)Bx$ if the Opponent concedes $q : (\forall x : A)Bx$. The core of the winning strategy is based on the fact that the Proponent can choose for the resolution of the instruction for the first component of the existential the same play-object that resolves the instruction of the first component of the universal posited by **O**. The explicit formulation of this process amounts to the Proponent making use of the equality $L^{\exists}(p) = L^{\forall}(q)$. Now, since the resolution of $L^{\forall}(q)$ will spread to $\mathrm{B}(L^{\forall}(q))$, we will have as a result that $\mathrm{B}(L^{\forall}(q))$ and $\mathrm{B}(L^{\exists}(p))$ are equal propositions.[36]

[36]One non negligible result of the interactive roots of definitional equality is that it provides a new insight into the dialogical take on the CTT approach to the notion of *harmony* as developed in Rahman and Redmond (2015a). Indeed, since the CTT approach to harmony, as mentioned above, is based on coordinating the elimination and introduction rule by means of definitional equality, and the latter, according to our analysis, corresponds to the strategic use of the formal rule, it follows that CTT- harmony is based on the strategic use of copy-cat interaction. Moreover, since, as argued in Rahman and Redmond (2015a), harmony in general can be achieved by the

The task ahead is to formulate rules that implement this explicitation-process as part of the development of a play. In the meanwhile let us display all the rules that determine explicit identity-expressions:

(Reflexivity within set)

Posit	Challenge	Defence
.		
X! $A : set$	**Y**$?_{set}$- *refl*	**X**! $A = A : set$

(Symmetry within set)

Posit	Challenge	Defence
.		
X! $A = B : set$	**Y**$?_B$- *symm*	**X**! $B = A : set$

(Transitivity within set)

Posit	Challenge	Defence
.		
X! $A = B : set$	**Y**$?_A$- *trans*	**X**! $A = C : set$
X! $B = C : set$		

(Reflexivity within A)

Posit	Challenge	Defence
.		
X! $a : \mathbf{A}$	**Y**? *a refl*	**X**! $a = a : A$

more fundamental notion of player-independence, the present analysis stresses the contribution of the dialogical theory of meaning that allows distinguishing the strategic use of definitional equality from a more basic notion of harmony. Perhaps we should speak of two different notions of harmony, one of them strategic (based on copy-cat plus definitional equality) and a semantic one (based on player-independence).

(Symmetry within A)

Posit	Challenge	Defence
.		
X! $a = b : A$	**Y**-$?_b$- *symm*	**X**! $b = a : A$

(Transitivity within A)

Posit	Challenge	Defence
.		
X! $a = b : A$	**Y**-I	**X**! $a = c : A$
X! $b = c : A$		

(Set-equality/Extensionality)

Posit	Challenge	Defence
X! $A = B$: *set*	**Y**-$?_{ext}$- $a : A$	**X**-!- $a : B$
	Y-$?_{ext}$-$a = b : A$	**X**! $a = b : B$

(Set-substitution)

Posit	Challenge	Defence
X! $B(x)$: *set* $(x : A)$	**Y**! $x = a : A$	**X**! $B(x/a)$: *set*
X! $B(x)$: *set* $(x : A)$	**Y**! $a = c : A$	**X**! $B(a)$=$B(c)$: *set*
X! $b(x) : B(x)$ $(x : A)$	**Y**! $a : A$	**X**! $b(a) : B(a)$
X! $b(x) : B(x)$ $(x : A)$	**Y**! $a = c : A$	**X**! $b(a)$=$b(c) : B(a)$

In these last rules, we have considered the simpler case where there is only one assumption in the proviso or context. The rules can obviously be generalized for provisos featuring multiple assumptions.

This ends the presentation of the dialogical notion of play-object and of the rules which give an abstract description of the local proceeding of dialogical games. Next we consider the global conditions taking part in the development of dialogical plays.

3.2.4 The Development of a Play

We will deal in this section with the other kind of dialogical rules called structural rules. These rules govern the way plays globally proceed and are therefore an important aspect of dialogical semantics. We will work with the following structural rules:

SR0 (Starting rule). Any dialogue starts with the Opponent positing initial concessions, if any, and the Proponent positing the thesis. After that the players each choose a positive integer called repetition ranks.

SR1i (Intuitionisitic Development rule). Players move alternately. After the repetition ranks have been chosen, each move is a challenge or a defence in reaction to a previous move and in accordance with the particle rules. The repetition rank of a player bounds the number of challenges he can play in reaction to a same move. Players can answer only against the *last non-answered* challenge by the adversary.[37]

SR2 ("Priority to formation" rule). ***O*** starts by challenging the thesis with the request '$?_{prop}$'. The game then proceeds by applying the formation rules first so as to check that the thesis is indeed a proposition. After that the Opponent is free to use the other local rules insofar as the other structural rules allow it.

SR3 (Modified Formal rule). ***O***'s elementary sentences cannot be challenged. However, ***O*** can challenge a ***P***-elementary move provided she did not herself play it before.

Since we have particle rules for elementary sentences involving the defence "*sic*(*n*)" we have no need for a formal rule which entitles a player to copy-cat some moves of the adversary.[38] We must however also ensure that the strictly internal aspect related to the idea of Geltung in the dialogical approach to meaning is not lost, and that the asymmetry between the player ***P*** who brings forward the thesis and his adversary ***O*** is accounted for. This is why the standard formal rule is replaced by this modified version.

SR4.1 (Resolution of instructions). Whenever a player posits a move in which instructions $I_1, \ldots, I_n$ occur, the other player can ask him to replace these instructions (or some of them) by suitable play-objects.

If the instruction (or list of instructions) occurs at the right of the colon and the posit is the tail of an universally quantified sentence or of an implication (so that these instructions occur at the left of the colon in the posit of the head of the implication), then it is the challenger who can choose the play-object. In these cases the player who challenges the instruction is also the challenger of the universal quantifier and/or of the implication.

Otherwise it is the defender of the instructions who chooses the suitable play-object. That is:

[37]This last clause is known as the *Last Duty First* condition, and is the clause making dialogical games suitable for Intuitionistic Logic, hence the name of this rule.

[38]But let us insist once more on the important point we raised in Sect. 3.2.3 contrary to standard dialogical games, copy-cat does not apply only to elementary sentences but to posits in which such sentences are associated with play-objects.

Posit	Challenge	Defence
X $\pi(I_1, \ldots, I_n)$	**Y** $I_1, \ldots, I_m$/? ($m \leq n$)	**X** $\pi(b_1, \ldots, b_m)$
		– if the instruction occurring at the right of the colon is the tail of either a universal or an implication (such that $I_1, \ldots, I_n$ also occur at the left of the colon in the posit of the head), then **$b_1, \ldots, b_m$ are chosen by the challenger**
		– Otherwise t**he defender chooses**

Important Remark In the case of embedded instructions $I_1(\ldots(I_k)\ldots)$, the substitutions are thought of as being carried out from I_k to I_1: first substitute I_k with some play-object b_k, then $I_{k\text{-}1}(b_k)$ with $b_{k\text{-}1}$ etc. until $I_1(b_2)$. If such a progressive substitution has already been carried out once, a player can then replace $I_1(\ldots(I_k)\ldots)$ directly.

SR4.2 (Substitution of instructions). During the play, when the play-object b has been chosen by any of the two players for an instruction I, and player ***X*** makes any posit $\pi(I)$, then the other player can ask to substitute I with b in this posit.

Posit	Challenge	Defence
X $\pi(I)$	**Y**? I/b	**X** $\pi(b)$
(where I/b has been previously established)		

The idea is that the resolution of an instruction yields a certain play-object for some substitution term, and therefore the same play-object can be assumed to result from any other occurrence of the same substitution term: instructions, after all, are functions and must yield as such the same play-object for the same substitution term.

SR5 (Winning rule for plays). For any p, a player who posits "$p : \bot$" looses the current play. Otherwise the player who makes the last move in a dialogue wins it.

In comparison to the rules of standard dialogical games, some additions in the rules we just gave have been made, namely SR2 and SR4.1-2. Also, the formal rule (here SR3) and the winning rule are a bit different. Since we made explicit the use of $\bot$ in our games, we need to add a rule for it: the point is that positing *falsum* leads to immediate loss. We could say that it amounts to a withdrawal.[39] Hence the formulation of the winning rule for plays above.

We need the rules SR4.1 and SR4.2 because of some features of CTT's explicit language. In CTT it is possible to account for questions of dependency, scope, etc. directly at the language level. In this way various puzzles, such as anaphora, get a convincing and successful treatment. The typical example, considered below, is the so-called donkey sentence "Every man who owns a donkey beats it". The two rules

[39] See Keiff (2007).

account for the way play-objects can be ascribed to what we have called instructions. See the dialogue in Sect. 3.4.4 for an application.

The rule SR2 is consistent with the common CTT practice to start demonstrations by checking or establishing the formation of propositions before proving their truth. Notice that this step also covers the formation of sets – membership, generation of elements, etc. – occurring in hypothetical posits and in quantifiers. In the current study, however, we can overlook this rule since we have restricted this work to the valid fragment of CTT: we can take it for granted that expressions are well formed. We will therefore only consider cases for which it is not necessary to carry out the formation steps since even if they were carried out, the players would always be able to justify that their expressions are well formed. We will, for this, always take examples guaranteeing, by the hypotheses introduced as initial concessions by the Opponent at the beginning of the play, that the expressions used are well formed.

What is more, it seems like we could liberalise the rule SR2. But because of the number of rules we have introduced, verifying this carefully is a delicate task that we will not carry out in this study. Let us for now simply mention that it seems sensible enough in dialogues to combine more freely the process linked to the formation rules with the development of a play. It does in fact seem perfectly consistent with actual practices of questioning the status of expressions introduced in the course of the game. Suppose for example that player ***P*** has posited '$p : \varphi \vee \psi$'. As soon as he has posited that the disjunction is a proposition – i.e., as soon as he has posited '$\varphi \vee \psi$: prop' – the other player knows how to challenge the disjunction and should be free to either keep on exploring the formation of the expression or to challenge the first posit. The point is that in a way it makes more sense to check whether φ is a proposition or not once (or if) ***X*** posits it to defend the disjunction. Doing so in a 'monological' framework such as CTT would probably bring various confusions, but the dialogical approach to meaning should quite naturally allow this additional dynamic aspect. Nonetheless, in order to generalise the equivalence result we have investigated here beyond the valid fragment of CTT (the reason why we have introduced rule SR2), it seems sensible in our view to clearly distinguish in a fashion close to CTT the steps linked to the formation from the other aspects of meaning.

The definitions of plays, games and strategies are the same as those given in appendix. Let us now recall them. A play for φ is a sequence of moves in which φ is the thesis posited by the Proponent and which complies with the game rules. The dialogical game for φ is the set of all possible plays for φ and its extensive form is nothing but its tree representation. Thus, every path in this tree which starts with the root is the linear representation of a play in the dialogical game at stake.

We say that a play for φ is terminal when there is no further move allowed for the player whose turn it is to play. A strategy for player ***X*** in a given dialogical game is a function which assigns a legal ***X***-move to each non terminal play where it is ***X***'s turn to move. When the strategy is a winning strategy for ***X,*** the application of the function turns those plays into terminal plays won by ***X.*** It is common practice to consider in an equivalent way an ***X***-strategy **s** as the set of terminal plays resulting when ***X*** plays according to **s**. The extensive form of **s** is then the tree representation of that set. For more explanations on these notions, see Clerbout (2014b). The

equivalence result between dialogical games and CTT is established by procedures of translation between extensive forms of winning strategies.

We have explained that the view of propositions as sets of winning strategies overlooks the level of plays and that an account more faithful to the dialogical approach to meaning is that of propositions as sets of play-objects. But play-objects are not the dialogical counterparts of CTT proof-objects, and thus are not enough to establish the connection between the dialogical and the CTT approach.

The local rules of our games – that is, the formation rules together with the particle rules – exhibit some resemblances to the CTT rules, especially if we read the dialogical rules backwards. But in spite of the resemblances, play-objects are in fact very different from CTT proof-objects. The case where the difference is obvious is implication – and thus universal quantification, which is similar. In the CTT approach, a proof-object for an implication is a lambda-abstract, and a proof-object of the tail of the implication is obtained by applying the function to the proof-object of the head. But in our account with play-objects, nothing requires that the play-object for the right-hand part is obtained by the application of some function.

From this simple observation it is clear that the connection between our games and CTT is not to be found at the level of plays. In fact it is well known that the connection between dialogues and proofs is to be found at the level of strategies (see, for example, (Rahman et al. 2009) for a discussion in relation to natural deduction). Even without the question of the relation to CTT, the task of describing and explaining the level of strategies is required, since it is a proper and important level of meaning analysis in the dialogical framework. This work has been developed in a recent volume (Clerbout and Rahman 2015), where a precise algorithm has been described that leads from winning strategies to CTT-demonstrations and back.

Summing up, we have play-objects which carry the interactive aspects of meaning-explanations. A proposition is the set of all possible play-objects for it, and a strategy in a game about this proposition is some subset of play-objects for it.

3.3 The Dialogical Take on the Axiom of Choice

In the present section we confront Per Martin- Löf's analysis of the Axiom of Choice with Jaakko Hintikka's (1996a) views on this axiom, who, to the best of our knowledge, was the first to provide a game-theoretical interpretation of it. Hintikka claims that his Game-Theoretical semantics (GTS) for Independence Friendly Logic justifies Zermelo's Axiom of Choice in a first-order way perfectly acceptable for the epistemic perspective of the constructivists. In fact, as pointed out by Jovanovic (2015) Martin-Löf's results lead to the following considerations:

1. Hintikka's preferred version of the Axiom of Choice is indeed acceptable for the constructivists and its meaning does not involve higher order logic.
2. However, the version acceptable for intuitionists is based on an intensional take on functions. From the point of view of Martin-Löf's (2006) intuitionistic approach, extensionality is the heart of the classical understanding of Zermelo's axiom and this is the real reason behind the rejection of it.

3. More generally, dependence and independence features that motivate IF-Logic, can be formulated within the frame of constructive type theory (CTT) without paying the price of a system that is neither axiomatizable nor has an underlying theory of inference – logic is about inference after all.

In this context it should be mentioned that very recent publications show that even in the frame of an extensional but constructive understanding of type theory the Axiom of Choice obtains (Sterling 2015). According to this perspective, the blame on Zermelo's axiom comes not from extensionality but from the unjustified assumption that the functionality of the relation in the antecedent, defined for specific setoids (roughly, extensional sets with a defined equivalence relation) yields a function that ranges over arbitrary quotient sets. It is this, so to say, over-generalization that is equivalent to the assumption of the third excluded. Hence, once more, the results show that there is no way to defend the evidence of Zermelo's axiom (we mean: its apparent logical validity) and defend at the same time third excluded. In the following sections we will develop the intuitionistic and intensional perspective of Martin-Löf rather than the constructivist extensional one. The latter deserves a thorough separate discussion that will not be deployed in the present paper.

3.3.1 Two Plays on the Axiom of Choice

Since the work of Martin-Löf (1984, pp. 50–51) the intensional formulation of the Axiom of Choice is *evident* in the sense that is logically valid. As pointed out by Bell (2009, p. 206) its logical validity entitles us to call it an *axiom* rather than a *postulate* (as in its classical or *extensional* version, that is not valid).[40] Jovanovic (2015) showed that, if we were to make explicit the domain and codomain of the function at the object language level, Hintikka's own formulation amounts to the following one – which for Hintikka's dismay is the intensional version of the AC as brought forward by Martin-Lof[41]:

$$(\forall x : A)\,(\exists y : B(x))\;\; C\,(x, y) \rightarrow (\exists f : (\forall x : A)\,B(x))\,(\forall x : A)\,C\,(x, f(x))$$

Before developing exhaustively the winning strategy for the intensional Axiom of Choice let us formulate the idea behind the dialogical approach by emulating Martin-Löf's (1984, p. 50)[42] own presentation of the informal constructive demonstration of it.

[40]Extensionality can be also rendered, provided uniqueness of the function, for a dialogical reconstruction of the proof see Clerbout and Rahman (2015).

[41]However, as Jovanovic (2015) discusses, Hintikka tries to render its meaning via a non-constructive semantics based on IF-logic.

[42]See too Bell (2009, p. 203–204) who makes use of the notation of Tait (1994) that is very close to that of the instructions of the dialogical frame, provided they occur in the core of strategy – that is, when they occur in those expressions that constitute a winning strategy. Indeed, Tait's functions π and π' corresponds to our left and right instructions – though we differentiate instructions for each logical constant adding an exponential to identify them. However, we do no have explicitly the

From the dialogical point of view the point is that **P** can copy-cat **O**'s choice for *y* in the antecedent for his defence of *f(x)* in the consequent since both are equal objects of type *B(x)*, for any *x : A*. Thus, a winning strategy for the implication follows simply from the meaning of the antecedent. This meaning is defined by the dependences generated by the interaction of choices involving the embedding of an existential quantifier in a universal one:

- Let us assume that the Opponent launches an attack on the implication and accordingly posits its antecedent – the play object for the antecedent being $L^{\rightarrow}(p)$. Let us further assume that with her challenge **O** resolves the instruction $L^{\rightarrow}(p)$, by choosing *v*.
- Then for any *x : A* chosen by **P**, there must be a play-object for the right component of *v* ($R^{\forall}(v)$, occurring in the antecedent.
- However, the play-object $R^{\forall}(v)$ (the right component of *v*) is a play-object for an existential and is thus composed by two play-objects such that the first one ($L^{\exists}(R^{\forall}(v))$, for any *x : A* is of type *B(x)*. and its right component, is, for any *x : A*, of type $C(x, L^{\exists}(R^{\forall}(v))$.
- Now, let **P** choose precisely the same play-object *v* for his defence of the existential in the consequent – the play-object for the consequent being $R^{\rightarrow}(p)$. Accordingly, the left play-object for the existential in the consequent is, for any *a : A*, of type *B(x)*. Thus, the left component of the play-object for the existential in the consequent is of the same type as the left component of the existential in the antecedent. Moreover, since **P** copies (while defending the existential) the choice of **O** (while resolving $R^{\rightarrow}(p)$) – namely *v* – we are entitled to say that the left component of the play-object for the existential in the consequent is exactly the same in *B(x)* as the left component of the existential occurring in the antecedent – i.e. $y = v(x) : B(x)$.
- Now, since in the antecedent *y* in *C(x,y)* is of type *B(x)*, for any *x : A*, and since, as already mentioned, *y* is equal to *v(x)* in *B(x)*, then it follows that *C(x,y)* in the antecedent is, for any *x : A*, intensionally equal to *C(x, v(x))* in the type *set*. More generally, and independently of **O**'s particular choice for the play-object for the antecedent, and independently of **O**'s particular choice of *x*, *C(x, y)* and *C(x, f(x))* are two equal sets (for any *x : A* and for *y : B(x)*).

From the two last steps it follows that **P** can copy-cat the play-object for the antecedent into the play-object for the consequent. This is the idea underlying a winning strategy for the Proponent for the Axiom of Choice. Also, these play-objects for antecedent and consequent are the ones relevant for the demonstration: one can then say that they are proof-objects. We will only deploy the plays that have been extracted of the extensive tree of all the plays. These plays constitute the so-called core of the strategy (that is, of the dialogical proof),[43] and they are triggered by the Opponent's options at move 9 when challenging the existential

function σ of Tait, though the result of the substitution of an instruction with a pair of embedded instructions – what we call it's resolution – will yield the pair of its components.

[43]For the process of their extraction and for the proof that these plays render the corresponding CTT demonstration see Clerbout and Rahman (2015).

posited by the Proponent at move 8. Since **O**'s repetition rank is 1, she cannot perform both challenges within one and the same play, hence the distinction between the following two plays. The first play corresponds in the demonstration to the introduction of the universal in the consequent, under the assumption of the antecedent. The second play develops all the points of the informal demonstration described above:

First play: Opponent's 9th move asks for the left play-object for the existential quantification on *f*

	O			P	
	H1*: C(x, y) : set (x : A, y : B(x))*			$p : (\forall x : A)\ (\exists y : B(x))\ C(x,y) \rightarrow (\exists f : (\forall x : A)\ B(x))\ (\forall x : A)\ C(x, f(x))$	0
	H2: *B(x) : set (x : A)*				
1	m:= 1			n:= 2	2
3	$L^{\rightarrow}(p) : (\forall x : A)\ (\exists y : B(x))\ C(x,y)$	0		$R^{\rightarrow}(p) : (\exists f : (\forall x : A)\ B(x))\ (\forall x : A)\ C(x, f(x))$	6
5	$v : (\forall x : A)\ (\exists y : B(x))\ C(x,y)$		3	$L^{\rightarrow}(p)/?$	4
7	$R^{\rightarrow}(p)/?$	6		$(v, r) : (\exists f : (\forall x : A)\ B(x))\ (\forall x : A)\ C(x, f(x))$	8
9	$?_L$	8		$L^{\exists}(v, r) : (\forall x : A)\ B(x)$	10
11	$L^{\exists}(v, r)/?$	10		$v : (\forall x : A)\ B(x)$	12
13	$L^{\forall}(v) : A$	12		$R^{\forall}(v) : B(w)$	26
15	$w : A$		13	$L^{\forall}(v) :/?$	14
19	$R^{\forall}(v) : (\exists y : B(w))\ C(w, y)$		5	$L^{\forall}(v) : A$	16
17	$L^{\forall}(v)/?$	16		$w : A$	18
21	$(t_1, t_2) : (\exists y : B(w))\ C(w, y)$		19	$R^{\forall}(v)/?$	20
23	$L^{\exists}((t_1, t_2) : B(w)$		21	$?_L$	22
25	$t_1 : B(w)$		23	$L^{\exists}(t_1, t_2)/?$	24
27	$R^{\forall}(v)/?$	26		$t_1 : B(w)$	28

Description

Move 3: After setting the thesis and establishing the repetition ranks **O** launches an attack on material implication.

Move 4: P launches a counterattack and asks for the play-object that corresponds to $L^{\rightarrow}(p)$.

Moves 5, 6: O responds to the challenge of 4. **P** posits the right component of the material implication.

Moves 7, 8: O asks for the play-object that corresponds to $R^{\rightarrow}(p)$. **P** responds to the challenge by choosing the pair (v, r) where v is the play-object chosen to substitute the variable f and r the play-object for the right component of the existential.

Move 9: O has here the choice to ask for the left or the right component of the existential. The present play describes the development of the play triggered by the left choice.

Moves 10–26: follow from a straightforward application of the dialogical rules. Move 26 is an answer to move 13, which **P** makes after he gathered the information the application of the copy-cat method (enclosed in the formal rule) requires.

Move 27–28: **O** asks for the play-object that corresponds to the instruction posited by **P** at move 26 and **P** answers and **win**s by applying copy-cat to **O**'s move 25. Notice that 28 this is not a case of function substitution: it is simply the resolution of an instruction.

Second play: Opponent's 9th move asks for the right play-object for the existential quantification on *f*

	O			P	
	H1*: C(x, y) : set (x : A, y : B(x))*			*p : (∀x : A) (∃y : B(x)) C(x,y) → (∃f : (∀x : A) B(x)) (∀x : A) C(x, f(x))*	0
	H2: *B(x) : set (x : A)*				
1	m:= 1			n:= 2	2
3	$L^{\rightarrow}$*(p) : (∀x : A) (∃y : B(x)) C(x, y)*	0		$R^{\rightarrow}$*(p) : (∃f : (∀x : A) B(x)) (∀x : A) C(x,f(x))*	6
5	*v : (∀x : A) (∃y : B(x)) C(x, y)*		3	$L^{\rightarrow}$*(p)/?*	4
7	$R^{\rightarrow}$*(p) /?*	6		*(v, r) : (∃f : (∀x : A) B(x)) (∀x : A) C(x, f(x))*	8
9	$?_R$	8		$R^{\exists}$*(v, r) : (∀x : A) C(x,* $L^{\exists}$*(v, r)(x)*	10
11	$L^{\exists}$*(v, r) /?*	10		$R^{\exists}$*(v, r) : (∀x : A) C(x, v(x))*	12
13	$R^{\exists}$*(v, r) /?*	12		*r : (∀x : A) C(x, v(x))*	14
15	$L^{\forall}$*(r) : A*	14		$R^{\forall}$*(r) : C(x, v(w))*	32
17	*w : A*		15	$L^{\forall}$*(r) : /?*	16
21	$R^{\forall}$*(v) : (∃y : B(w)) C(w,y)*		5	$L^{\forall}$*(v) : A*	18
19	$L^{\forall}$*(v) /?*	18		*w : A*	20
23	*(*t_1*, *t_2*) : (∃y : B(w)) C(x,y)*		21	$R^{\forall}$*(v) /?*	22
25	$L^{\exists}$*((*t_1*, *t_2*): B(w)*		23	$?_L$	24
27	t_1*: B(w)*		25	$L^{\exists}$*(*t_1*, *t_2*) /?*	26
29	$R^{\exists}$*(*t_1*, *t_2*): C(w, *t_1*)*		23	$R_?$	28
31	t_2*: C(w, *t_1*)*		29	$R^{\exists}$*(*t_1*, *t_2*) /?*	30
33	$R^{\forall}$*(r) /?*	32		t_2*: C(w, v(w))*	34
35	*v(w) /?*	34		t_2*: C(w, *t_1*)*	42
				*< C(w, *t_1*) = C(w, *t_1*/ v(w)) : set >*	
41	*C(w, *t_1*)= C(w, *t_1*/ v(w)) : set*		H1$?_{subst}$	*v(w) =* t_1*: B(w)*	36
37	*v(w) =* t_1*: B(w) ?*	36		*sic (39)*	40
39	*v(w) =* t_1*: B(w)*		5, 18, 21, 25	$?_{\Pi\text{-}eq}$.	38

Description

Move 9: Until move 9 this play is the same as the previous one. In the present play, in move 9 the Opponent chooses to ask for the right-hand side of the existential posited by **P** at 8.

Moves 10–34: the Proponent substitutes the variable f by the instruction correspondent to the left-hand component of the existential, i.e., $L^{\exists}(v, r)$. By this **P** accounts for the dependence of the right-hand part on the left-hand component. The point is that the local meaning of the existential requires this dependence of the right component to the left component even if in this play the Opponent, due to the restriction on rank 1, can ask only for the right-hand part.

The conceptually interesting moves start with 35, where the Opponent asks **P** to substitute the function. As already pointed out, in order to respond to 35 the Opponent's move 31 is not enough. Indeed the Proponent needs also to posit $C(w, t_1) = C(w, t_1/v(w)) : set$. **P** forces **O** to concede this equality (41), on the basis of the substitutions w/x and t_1/y on H1 (we implemented the substitution directly in the answer of **O**) given the Π-*equality* $v(w) = t_1$ in $B(w)$ (36), and given that this Π-*equality* yields the required set equality. Moreover, **P**'s posit of the Π-*equality* (36) is established and defended by moves 38–40.

3.3.2 The Core of the Winning Strategy

The core of the winning strategy in the dialogical game for the Axiom of Choice consists of the two plays we have just described, written in a linear way and with a ramification when the Opponent can choose between asking for the left and asking for the left at her move 9.

This results in the tree-like structure given on next page. In order to identify the dialogical source of each move we make use of *[? n]* to indicate the attacked line and and *[m]* to indicate the challenge of player **X** that triggered the posited defence of **Y.**

O H1*: C(x, y) : set (x : A, y : B(x))*
O H2: *B(x) : set (x : A)*
0. **P** *p : (∀x : A) (∃y : B(x)) C(x,y) → (∃f : (∀x : A) B(x)) (∀x : A) C(x, f(x))*
1 **O** n : = 1
2. **P** m : = 1
3 **O** $L^{\rightarrow}(p)$ *: (∀x : A) (∃y : B(x)) C(x, y)* [? 0]
4 **P** $L^{\rightarrow}(p)$ */?* [? 3]
5 **O** *v : (∀x : A) (∃y : B(x)) C(x, y)* [4]
6 **P** $R^{\rightarrow}(p)$ *: (∃f : (∀x : A) B(x)) (∀x : A) C(x,f(x))* [3]
7 **O** $R^{\rightarrow}(p)$ */?* [? 6]
8 **P** *(v, r) : (∃f : (∀x : A) B(x)) (∀x : A) C(x, f(x))* [7]

9 **O** $?_L$ [? 8]
10 **P** $L^{\exists}(v, r)$ *: (∀x : A) B(x)*[9]
11 **O** $L^{\exists}(v, r)$ */?* [? 10]
12 **P** *v : (∀x : A) B(x)* [11]
13 **O** $L^{\forall}(v)$ *: A* [? 12]
14 **P** $L^{\forall}(v)$ *: /?* [? 13]
15 **O** *w : A* [14]
16 **P** $L^{\forall}(v)$ *: A* [? 5]
17 **O** $L^{\forall}(v)$ */?* [? 16]
18 **P** *w : A* [17]
19 **O** $R^{\forall}(v)$ *: (∃y : B(w)) C(w, y)* [16]
20 **P** $R^{\forall}(v)$ */?*[? 19]
21 **O** (t_1, t_2) *: (∃y : B(w)) C(w, y)* [20]
22 **P** $L_?$ [? 21]
23 **O** $L^{\exists}((t_1, t_2)$ *B(w)* [22]
24 **P** $L^{\exists}(t_1, t_2)$ */?* [? 23]
25 **O** t_1 *: B(w)* [24]
26 **P** $R^{\forall}(v)$ *: B(w)* [13]
27 **O** $R^{\forall}(v)$ */?* [? 26]
28 **P** t_1 *: B(w)*[27]

9 **O** $?_R$ [? 8]
10 **P** $R^{\exists}(v, r)$ *: (∀x : A) C(x,* $L^{\exists}(v, r)(x)$ [9]
11 **O** $L^{\exists}(v, r)$ */?* [? 10]
12 **P** $R^{\exists}(v, r)$ *: (∀x : A) C(x, v(x))* [11]
13 **O** $R^{\exists}(v, r)$ */?* [? 12]
14 **P** *r : (∀x : A) C(x, v(x))* [13]
15 **O** $L^{\forall}(r)$ *: A* [? 14]
16 **P** $L^{\forall}(r)$ *: /?* [? 15]
17 **O** *w : A* [16]
18 **P** $L^{\forall}(v)$ *: A* [? 5]
19 **O** $L^{\forall}(v)$ */?* [? 18]
20 **P** *w: A* [19]
21 **O** $R^{\forall}(v)$ *: (∃y : B(w)) C(w,y)* [18]
22 **P** $R^{\forall}(v)$ */?* [? 21]
23 **O** (t_1, t_2) *: (∃y : B(w)) C(w ,y)* [22]
24 **P** $?_L$ [? 23]
25 **O** $L^{\exists}((t_1, t_2)$ *: B(w)*[24]
26 **P** $L^{\exists}(t_1, t_2)$ */?* [? 25]
27 **O** t_1 *: B(w)* [26]
28 **P** $?_R$ [? 23]
29 **O** $R^{\exists}(t_1, t_2)$*: C(w,* t_1*)* [28]
30 **P** $R^{\exists}(t_1, t_2)$ */?* [? 29]
31 **O** t_2 *: C(w,* t_1*)* [30]
32 **P** $R^{\forall}(r)$ *: C(w, v(w))* [15]
33 **O** $R^{\forall}(r)$ */?* [? 32]
34 **P** t_2 *: C(w, v(w))* [33]
35 **O** *v(w) /?* [? 34]
36 **P** *v(w) =* t_1 *: B(w)* [? H1]
37 **O** *v(w) =* t_1 *?* [? 36]
38 **P** $?_{\Pi\text{-}eq}$ [? , 5, 18, 21, 25]
39 **O** *v(w) =* t_1 *: B(w)* [38]
40 **P** *sic* [39]
41 **O** *C(w,* t_1*) = C(w,* t_1 */ v(w)) : set* [36]
42 **P** t_2 *: C(w,* t_1*) < C(w,* t_1 */ v(w)) : set* > [35]

3.3.3 Final Remarks

According to Hintikka, the following formulation of the Axiom of Choice,

$$\forall x \exists y\, C(x, y) \rightarrow \exists f \forall x\, C(x, f(x))$$

(where it is left implicit that $\forall x$ quantifies over, say the set A, $\exists y$ quantifies over, say the set B, and $\exists f$, over the set $(\forall x : A)\, Bx$)

is perfectly acceptable for the constructivists. Let us recall, that the GTS reading of its truth amounts to the existence of a winning strategy for Eloise in a game G ($\forall x \exists y\, C(x,y)$). The latter amounts to finding a "witness individual" y dependent upon x, such that $C(x,y)$ is true – notice how close this formulation is to Martin-Löfs "informal" description of the proof of the axiom. In other words, the existence of a winning strategy for that game provides *a proof that the proposition S(x,y) is true in the model* (Hintikka 1996a, ch 2). Hintikka claims that it is the GTS reading that makes the Axiom of Choice acceptable for the constructivists:

> *Moreover, the rules of semantical games should likewise be acceptable to a constructivist. In order to verify an existential sentence* $\exists x S[x]$ *I have to find an individual b such that I can (win in the game played with) S[b]. What could be a more constructivistic requirement than that? Likewise, in the verification game* $G(S_1 \vee S_2)$ *connected with a disjunction* $(S_1 \vee S_2)$*, the verifier must choose* S_1 *or* S_2 *such that the game connected with it (i.e.,* $G(S_1)$ *or* $G(S_2)$*) can be won by the verifier. Again, there does not seem to be anything here to alienate a constructivist.* (Hintikka 1996a, p. 212)

As mentioned above Hintikka is right, this is acceptable for the constructivists, but the reason is the underlying intensionality of the choice function and this assumes an underlying intuitionistic and not a classical logic as Hintikka was aiming at. An alternative option is to formulate the constructivist version of Axiom of Choice in an IF-setting. In such a setting the rejection of the constructivists is rendered as the rejection of the de re occurrence of the choice function in the consequent. The price to pay is known: the resulting formal system is not axiomatizable. This is a too high price, given that its truth can be made evident with the means of CTT.[44] Moreover, the IF-reconstruction of the version of the Axiom of Choice rejected by the constructivist is not equivalent to the third-excluded. Hence, the IF-reconstruction is really changing the subject at stake.

Hintikka's intention was to offer a realist foundation of mathematics on a first-order level in a way that all classical mathematics can be comprised and that it can still be acceptable for a constructivist. As pointed out by Göran Sundhom (2016, forthcoming) under constructivist reading IF logic is granted to be a first-order logic, but in that case not all of the classical mathematics can be saved. However, as we think we have shown in this section, a game-theoretical interpretation that meets the epistemic requirements of the constructivists is possible, but this produces a dialogical version with an antirealist approach to meaning rather than a GTS

[44] Cf. Jovanovic (2013).

interpretation with an underlying model of formal semantics. In fact, the point of Hintikka is still a valuable one, the winning strategy of a universal quantifier is a function, such that, if there is an existential embedded in the universal expression; the verifier will make his choices (for the existential) dependent upon the ones of the falsifier. This is what the constructive reading is about. Moreover, Hintikka's point can be carried out precisely in a language that allows those *dependent proof-objects* (i.e. functions) to be expressed at the object language level of first-order logic: the functions are in fact nothing more than the truth-makers (proof-objects/winning-strategies) of the corresponding first-order expressions. This point is crucial for the understanding of anaphora.

3.4 GTS and Dialogical Logic on Anaphora

Hintikka's and associates' work on anaphora, based on Game-Theoretical semantics (GTS), constitutes a landmark in the field and it triggered many valuable contributions and discussions. The landmark-setting of Sundholm (1986) and the further developments of Ranta (1994) show that CTT has the means to provide a precise analysis of anaphoric expressions. In this section we will compare the GTS approach to anaphora with that of the dialogical approach. It is our opinion that GTS approach provides – from the viewpoint of its use in natural language – an understanding of anaphora that is very close to actual linguistic practice: recall that anaphora is one of the main structures of conversational contexts. Thus, a semantics based on interaction seems to be indeed the most suitable approach. However, according to our view, the extension of the dialogical framework discussed in the precedent sections contains both the contentual (first-order) features of CTT placed at the object language level and the interactive aspects of GTS.

3.4.1 The GTS Approach to Anaphora

The issue is to find a satisfying semantic analysis of anaphoric expressions occurring in sentences such as:

1. *If Michael smiles he is happy*.
2. *If a man smiles he is happy*.
3. *Every man that smiles is happy*;

and of more problematic examples such as the famous donkey sentence:

4. *Every man who owns a donkey beats it*.

Texts such as

5. *Nick stood up. He was all right. He looked up at the lights of the caboose going out of sight around the curve.*

Conversations such as

6. *Bernadette drinks her coffee with 5 cubes of sugar. Bernadette who? Ah, the Ivorian doctorate student of Shahid Rahman. She will get sick eventually.*

The first sentence apparently is not problematic. The pronoun "he" has a strict interpretation (Michael), so it can be treated as a singular term. The issue is to provide a satisfying semantic analysis of pronouns "he" or "it" which becomes more challenging when there is interplay between pronouns and indefinites such as in the other cases.

Since the main aim of the present paper is to motivate studies on the interface between games and CTT we will not really delve into all the fine subtleties of all the different anaphora-cases but rather centre our attention to the general case and then the example of the donkey sentence. Moreover, since, as mentioned above, we are convinced that the interaction aspect stressed by Hintikka's analysis is crucial to the understanding of anaphora we will start with a brief overview of Hintikka's et alii approach.[45]

According to Hintikka's analysis, if a quantifier is understood as a logical expression then we are speaking of its *priority scope* in relation to the rest of the sentence, but if it is understood as the antecedent for anaphoric pronoun that appears in the rest of the sentence then we are dealing with its *binding scope*. It is a pity, from Hintikka's point of view, that those two different moments are expressed by the same syntactic expression. At first glance, it is appealing to interpret anaphoric pronouns as variables available for quantification. But Hintikka contests this view:

> [...] *they do not behave like bound variables. An anaphoric pronoun does not receive its reference by sharing it with the quantifier phrase that is its "head", any more than a definite description does. An anaphoric pronoun is assigned a reference in a semantical game through a strategic choice of a value from the choice set by one of the players. When the member of the choice set whose selection is a part of the winning strategy of the player in question happens to be introduced to the choice set by a quantifier phrase, that phrase could perhaps be called the head of the pronoun. But, as was pointed out, the origin of the members of the choice set does not matter at all in the semantical rules for anaphoric pronouns.* (Hintikka, 1997, p. 530)

What Hintikka seems to be aiming at is to stress the cases where the relevant meaning-relation between a quantifier and the quantified expression is a relation of *dependence* rather than a compositional one, and that such dependences should be understood as interaction. This was one of the main motivation for a GTS-approach to anaphora and it was further developed by Sandu and his associates (Sandu 1997, Sandu and Jacot 2012). In other words, the main point of anaphora is dependence, dependence is interaction, and thus a semantics of interaction is required. Let us see the latter point before we study its applications.

[45] Already in 1985 Hintikka and Kulas used GTS in order to provide semantics of definite descriptions.

3.4.2 GTS

According to GTS, meaning is obtained through the interaction of two players, *Verifier* and *Falsifier*,[46] in a semantic game. The game starts from the whole formula and descending to the atomic formulas, the truth of which is checked in the model – in other words, the attribution of meaning goes exactly the inverse way of standard Tarksi-style semantics which proceeds "from inside out". This "outside-in" approach seems to be much more promising in the treatment of meaning and that is the feature that GTS shares with the dialogical approach we will discuss in Sect. 3.4.4.

The game is defined as follows:

– *Definition:*

Let Eloise and Abelard be the players in a game. Eloise is the initial verifier, trying to defend the sentence at stake and Abelard is the initial falsifier, trying to deny it.

A semantic game $G(\varphi)$ for the sentence φ begins with φ. The game is played in the model $\mathbb{M}$ with a given language L. Through various stages of the game, players will consider either the sentence φ or other sentence φ' obtained from φ during the development of the game. The game is played with well-defined rules.

R$\vee$ – disjunction-rule: $G(\varphi_1 \vee \varphi_2)$ starts by the choice of the player who has (in G) the role of verifier, for φ_i (i = 1 or 2). The game continues as $G(\varphi_i)$.

R$\wedge$ – conjunction-rule: $G(\varphi_1 \wedge \varphi_2)$ starts by the choice of the player who has (in G) the role of falsifier, for φ_i (i = 1 or 2). The game continues as $G(\varphi_i)$.

R$\exists$ – rule: $G(\exists x\, Sx)$ starts by the choice of the player who has (in G) the role of verifier, of one member from the domain of $\mathbb{M}$ for x. If the name of the individual is a, the game is played as $G(Sa)$.

R$\forall$ – rule: $G(\forall x Sx)$ starts by the choice of the player who has (in G) the role of falsifier, of one member from the domain of M for x. If the individual is a, the game is played as $G(Sa)$.

R$\neg$ – rule: $G(\neg\varphi)$ is played the same as $G(\varphi)$ except that players change their roles.

R – atomic-rule for the atomic sentences: if A is an atomic sentence that is true, the verifier wins. If the sentence is false the falsifier wins.

Each application of the rules eliminates one logical constant, so that in a finite number of steps eventually the rule for atomic sentences must be applied. Truth of an atomic sentence is determined by the model $\mathbb{M}$ with respect to which $G(\varphi)$ is played. In other words, the game G assumes the interpretation of all non-logical constants in the model $\mathbb{M}$, and it provides the (model-theoretical) meaning of the primitive symbols of a given interpreted first-order language.

[46] Sometimes called *Myself and Nature*.

Finally, here are the truth and falsity conditions for arbitrary formulae:

– *Definition:*

(a) φ is true in model $\mathbb{M}$ ($\mathbb{M} \vDash_t \varphi$) if and only if there is a winning strategy for Eloise in the game *G(φ)* played in $\mathbb{M}$.
(b) φ is false in model $\mathbb{M}$ ($\mathbb{M} \vDash_f \varphi$) if and only if there is a winning strategy for Abelard in the game *G(φ)* played in $\mathbb{M}$.

3.4.3 GTS, Anaphora and Branching Quantifiers

In the GTS frame, strategies of players that are introduced on the semantic level are Skolem functions that tell a player which disjunct/ conjunct or which individual in the model to choose every time it is her turn to play. Given the prenex normal form of a formula, we obtain its Skolem form by replacing systematically every existential quantifier by an appropriate Skolem function the argument of which is a variable bound by a universal quantifier in the scope of which that existential quantifier lays. Hintikka's main idea is that a Skolemization of the strategies yields a correct analysis of the anaphora. Sandu and Jacot (2012) added the further step of introducing the skolemization in the object language level by means of *Skolem terms*.

Let us see how this works in the case of the donkey sentence. The analysis starts with the GTS approach of the following universal:

Every man owns a donkey.

In a game played for this sentence it is the falsifier who first chooses an individual that satisfies the predicate of being a man. Then it is on the verifier to find a donkey owned by that individual in order to win the game. This game can be represented as a tree with branches that shows all possible outcomes of the game (for any individual chosen by falsifier). The strategy of the verifier is then a function f that for any individual a, chosen by falsifier gives as a result *f(a),* that is a donkey owned by a. The sentence given above is then formalised as

$$\forall x\Big(Man(x) \rightarrow Donkey\ \ (f(x) \wedge Own\ \ (x, f(x)))\Big).$$

We can now turn to the anaphoric pronoun in the sentence 4. A solution for the problematic anaphora is found with help of a Skolem term. The pronoun "it" is a copy of a Skolem term in the antecedent. The formalisation of 4 thus is:

$$\forall x\,(((Man(x) \wedge Donkey\ \ (f(x))) \wedge Own\ \ (x, f(x))) \rightarrow Beats\,(x, f(x)))\,.$$

Sandu and Jacot (2012, p. 620) claim that Skolem terms are very useful semantic tools for anaphora because they keep track of the entire history of a play of a game. All the variables bound by quantifiers superior to the indefinites are found as the arguments of each Skolem term. This solution combines at once the quasi-referential view on quantifiers, which is appealing when an anaphoric pronoun appears in a sentence, and the idea of semantic dependency, which is needed both when there is a nesting of indefinites and where there is an interplay of indefinites with quantifiers.

Moreover, this method can be extended to more complicated cases of dependence such as the ones of Henkin's branching quantifiers, that were before analysed by combining GTS with *Independence friendly first-order logic* (for short: IF).[47] Indeed, the IF-analysis of the case of branching quantifiers yields:

$$\left.\begin{matrix}\forall z \exists u \\ \forall x \exists y\end{matrix}\right\} \quad S(x, y, z, u)$$

cannot be expressed with one (linearly disposed) sentence of classical first-order logic. However this can be done in IF in the following way:

$$\forall x \forall z \ (\exists y / \forall z)(\exists u / \forall x) \ S(x, y, z, u)$$

where the slashes indicates that $\exists y$ *(*$\exists u$*)* is independent of $\forall z$ *(*$\forall x$*).*

In some cases, the slash does not contribute to anything that could not be expressed without it, but in others it allows to express structural features that we would not otherwise be able to express in standard first-order logic. Walkoe showed that the expressive power of formulas with branching quantifiers is precisely that of existential second-order logic (Walkoe 1970).[48] Independently, Walkoe and Enderton also showed that every existential second-order sentence $\sum^1_1$ is equivalent to second-order truth or falsity condition of an IF sentence. (Walkoe 1970; Enderton 1970). Thus, IF logic captures exactly the expressive power of Henkin's branching quantifiers, though according to Hintikka IF is first-order.[49]

A classic example of a natural language sentence that involves branching quantifiers from Hintikka (1973, p. 344) is:

[47]IF first-order logic is an extension of first-order logic, involving a specific syntactic device '/' (slash, independence indicator), which has at the object language level the same effect as the meta-level modifier 'but does not depend on'. IF was introduced by Jaakko Hintikka and Gabriel Sandu in their article 'Informational Independence as a Semantical Phenomenon' (1989); other early sources are Hintikka's booklet *Defining Truth, the Whole Truth, and Nothing but the Truth* (1997) and Sandu's Ph.D. thesis (1991).

[48]Existential second-order logic is a fragment of second-order logic that consists of a formula in the form $\exists x_1 \ldots \exists x_n \Psi$, where $\exists x_1 \ldots \exists x_n$ are second-order quantifiers and Ψ is a first-order formula.

[49]Feferman (2006) and Väänänen (2001) rose however the question whether IF logic is really first-order logic. Tulenheimo (2009) provides some elements to defend Hintikka's view. Curiously, Sundholm (2016, forthcoming) shows that those dependences and independences that motivate Hintikka's introduction of IF can be formulated in CTT first-order logic.

Some relative of each villager and some relative of each townsman hate each other.

If we formulate this in the new Skolem-terms-frame proposed by Sandu and Jacot (2012) we obtain:

$$\forall x \forall z\,((Villager(x) \wedge Townsman(z)) \rightarrow (Relative\ (x, f(x)) \wedge Relative\,(z, g(z)) \wedge Hate\,(f(x), g(z)))$$

Or quantifying over the functions:

$$\exists f \exists g \forall x \forall z\,((Villager(x) \wedge Townsman(z)) \rightarrow (Relative\ (x, f(x)) \wedge Relative\,(z, g(z)) \wedge Hate\,(f(x), g(z))))$$

According to Hintikka, IF allows us to take into account different patterns of dependency among logical expressions that can appear in a sentence, and it is thus more appropriate for the translation of natural languages than other approaches are.[50] Moreover, all this can be done at the first-order level. Similar can be said of the functional approach. Now, the first-order reading of the IF-formulation has been contested and the issue has not been settled yet. However, as already mentioned in the introduction and at the end of the section on the Axiom of Choice, the point is that the simple move of substituting Skolem functions by proof-objects of the quantified propositions under consideration yields a straightforward first-order reading: the functions at stake are the truth-makers of the propositions involving quantifier dependences. Furthermore the dialogical approach contributes to the game-theoretical approach by providing the elementary constituents of which the dependent proof-objects are made of. Indeed the dialogical approach provides the method to build the dependent proof-objects underlying a suitable winning strategy. We undertake this task in the next section.

3.4.4 The Dialogical Approach to Anaphora

In what follows we will give a dialogical account of anaphora making use of CTT. We will argue that the GTS approach, that puts the accent on expressing dependence relation in terms of choices resulting from interaction, is indeed a good way to

[50] Strategies for players in a game for a given sentence are expressed by existential second-order sentences, usually noted as $\sum^1_1$. According to Hintikka (1997, p. 523) *"this second-order statement expresses the logical form of the given natural-language sentence. It is equivalent to an IF first-order sentence, which can also be considered as the translation of the given natural language sentence into logical notation."*. The existential part of second-order logic exceeds in expressivity classical first-order logic, and since IF is equivalent to $\sum^1_1$, so does independent friendly logic (IFL).

deal with anaphora. However, our approach is closer to the recent Skolem-term framework developed by Sandu and Jacot than to the original analysis of Hintikka, though the dialogical framework can also deal with the more complicated cases involving branching quantifiers without making use of the formal system IF, and though, as mentioned above, we see Skolem-terms as the introduction into the object language level of dependent-proof-objects constituted by play-objects.

From the more general point of view of philosophy of language the dialogical approach to anaphora seems to match with a weakened version of Brandom's view on the relations between anaphora and deixis.[51] As pointed out by Penco (2005) the core of Brandom's strong claim for the conceptual priority of anaphora with respect to deixis – according to Brandom (1994, pp. 464–468) deixis presupposes anaphora – is based on the observation that the *capacity of pronouns to pick up a reference from an anaphoric antecedent is an essential condition of the capacity of other tokens (which can serve as such antecedents) to have references determined* (Penco 2005, p. 182). And the argument is somehow plausible, if we are thinking of *repeatable* referential situations. In such situations it is the anaphoric structure that allows us to re-identify what has been referred to by an indexical – Brandom ascribes the role of *anaphoric initiator* to indexicals. On this view, anaphoric initiators can trigger anaphoric chains. More precisely, according to Brandom, as soon as we use demonstratives and indexicals, we are beginning to keep track of an object via a possible anaphoric chain – this is, according to Brandom, the very point of the use of demonstratives and indexicals. However, as discussed by Penco (2005, pp. 182–184) indexicals do not only have the role of anaphoric initiators: they also perform the function of connecting general beliefs with contexts. Also, an indexical is sometimes used only once, that is to say: without initiating an anaphoric chain. Therefore, it seems that a more prudent way to express the point reduces to the observation that deixis and anaphora should be thought together. Still, this does not change really the core of Brandom's remark that the function of indexicals and demonstratives is not exhausted in their unrepeatable occurrence.

Be that as it may, the dialogical approach to pronouns takes them in their anaphoric role. Furthermore, since the dialogical approach to anaphora is based on the CTT-framework, also the dependence upon a context can be thought as having an anaphoric structure. In fact, Ranta (1994, p 78) introduces *pronominalisation rules for inference* in a CTT-frame in order to make explicit dependence of anaphoric pronouns upon the context.[52] But context is understood as the assumption that the picked object is of a given type (e.g. the assumption that x is of type A). Thus the pronouns (and more generally the indexicals) dependence upon a context is understood as a reference to any object of appropriate type.

For example, Ranta's inferential rules for the pronoun *he* and *she* deploy the identity mapping on the set of *man* and *she* as identity mapping on the set of *woman* as resulting for the contextual dependence:

[51] Though as discussed in the last paragraphs of the present section our analysis differs from the one of Brandom.

[52] Context should be understood in the technical sense of CTT.

$$\frac{a : man}{he(a) : man}$$

And the rule of a substitution[53]:

$$\frac{a : man}{he(a) = a : man}$$

It is important to notice that these rules do not really assume that an instance of the type given by the context is necessarily a constant expression, it could well be a variable (and then the context is in fact an open assumption). Moreover, it is possible to generalize the rule for embedded dependences, which renders Brandom's point on anaphoric chains. If we put all together the following inference rule described by Ranta (1994, p. 80) obtains:

$$\frac{a\,(x_1,\ldots,x_n) : man\;\; (x_1 : A_1,\ldots,x_n : A_n\,(x_1,\ldots,x_{n-1}))}{\begin{array}{c} he\,(a\,(x_1,\ldots,x_n)) : man\,(x_1 : A_1,\ldots,x_n : A_n\,(x_1,\ldots,x_{n-1}))\,, \\ he\,(a\,(x_1,\ldots,x_n)) = a\,(x_1,\ldots,x_n) : man\,(x_1 : A_1,\ldots,x_n : A_n\,(x_1,\ldots,x_{n-1})) \end{array}}$$

After the application of such rule (rules), it is possible to drop the argument a and the bare pronoun *he* can be used – in the context A. This can be formulated by an additional "sugaring" rule such as

$$he(a)\,(a : A) \lhd he$$

In the dialogical framework, Ranta's pronominalization rules are understood as the intertwining of commitments and entitlements that characterizes Brandom's overall view on meaning:

If player **X** posits that *he(a): man*, then his adversary can challenge this posit by asking him to show that a is of the type *man*. Since we would like to include variables, the best is to make use again of instructions,

Posit	Challenge	Defence
X $he(I^{pron})$*: man*	**Y** $?_{I^{pron}:man}$	**X** I^{pron}*: man*

The resolution of the instruction in I^{pron}*: man* allows the defender to introduce explicitly an identity within the set *man*:

[53] In Ranta (1994, p. 78) those two rules (identity mapping and substitution) are united in one rule with two conclusions.

Posit	Challenge	Defence
$\mathbf{X}\, I^{pron} : man$	$\mathbf{Y}_{Ipron,man/?:}$	$\mathbf{X}\, he(I^{pron}) = a : man$

A third rule implements the argument-dropping rule mentioned above: If a player brought forward an identity of the form described above, then the challenger can use this identity to substitute, say, he(a) for a, wherever he(a) occurs. In this rule we assume that instruction I^{pron} has already been substituted by a suitable play-object.

Posit	Challenge	Defence
$\mathbf{X}\, he(a) = a : man$	$\mathbf{Y}\ ?_{a/he(a)}$	$\mathbf{X}\, \varphi[a]$
...		
$\mathbf{X}\, \varphi[he(a)]$		

Because of the recursivity of the rules we will not write down explicitly the case of chains of dependences. Similar rules can be formulated for *she*. The case of *it* requires more care, since it's type might vary from context to context. Anyway this type-variation seems to apply to all pronouns (e.g. *she : ship*).

Let us come back to our example of the happy man:

If a man smiles he is happy.

The idea is that in order to obtain the interpretation for a pronoun *he*, we first formalize the first part "*A man smiles*" as:

$$(\exists x : man)\ \ smiles(x),$$

and then we consider the sentence "*he is happy*" in the context[54]

$$z : (\exists x : man)\ \ smiles(x).$$

This analysis yields the following formalisation:

$$(\forall z : (\exists x : man)\ \ smiles(x))\ \ happy\left(he\ \ \left(L^{\forall}(z)\right)\right)$$

According to the rules given in Sect. 3.2, the left part of the universal given above consists of the set of all men that smile and the right part claims that an object chosen from that set is happy. The left part of universal is $L^{\forall}(z)$, that is, the set of all men that smile: what the pronoun *he* does is to pick up one individual of the set of the smiley men (the set *(∃x: man) smiles(x)*). Let us deploy a play that illustrates both

[54] In Ranta (1994, p. 79) the example is "If a man walks he talks".

the analysis and use of the rules. For the sake of simplicity we do not make use of the instruction I^{pron} – after all the pronoun has already picked an instruction. We will also ignore the moves involving the choice of repetition ranks:

	O			P	
				(∀z: (∃x: man) smiles(x)) happy(he $(L^{\forall}(z))$	0
1	$L^{\forall}(z)$*: (∃x: man) smiles(x)*	*0*	1	$R^{\forall}(z)$ *: happy(he(*$L^{\forall}(z)$*))*	12
3	*a : (Ex : man) smiles(x)*			$L^{\forall}(z)$*/?*	2
5	$L^{E}(a)$ *: man*		3	*?L*	4
7	$R^{E}(a)$ *: smiles(*$L^{E}(a)$*)*		3	*?R*	6
9	a_1*: man*		5	$L^{E}(a)$*/?*	8
11	a_2*: smiles(*a_1*)*		7	$R^{E}(a)$*/?*	10
13	*?*$L^{\forall}(z)$*/? ...*	12		$R^{\forall}(z)$ *: happy(he(*a_1*))*	14
15	$?_{a1:\,man}$	14		a_1*: man*	16
17	$1^{he,man}/?_{:}$	16		$R^{\forall}(z)$ *: happy (he(*a_1*)=* a_1*)*	18
19	$?_{\,a/he(a)}$	18		$R^{\forall}(z)$ *: happy(*a_1*)*	20
21	*?*$R^{\forall}(z)$*/*	20		**P** loses unless he can force **O** to concede that there is a play-object *b* for *happy(*a_1*),* such that it allows **P** to choose *b* for $R^{\forall}(z)$ while responding to the challenge of move 21 on move 20	

Description

Move 0: **P** states the thesis.
Move 1: **O** challenges the universal by positing an arbitrary man that smiles, that is *z: (∃x: man) smiles(x).*
Move 2: **P** counterattacks by asking who that man is.
Move 3: **O** responds by choosing some play-object.
Move 4: Since *a* is a play-object for an existential, it is constituted by two parts: **P** starts by asking for its left part.
Move 5: **O** answers that *L(a)* is a man.
Move 6: **P** challenges now the right part of the existential.
Move 7: **O** responds to the attack.
Move 8: **P** asks **O** to resolve the instruction occurring in the expression brought forward in move 5
Move 9: **O** responds by choosing a_1.
Move 10: **P** asks **O** to resolve the instruction occurring in the expression brought forward in move 7.
Move 11: **O** responds by choosing a_2.
Move 12: **P** answers now the challenge of move 1**.**
Move 13: **O** asks **P** to resolve the instruction $L^{\forall}(z)$ occurring in the expression brought forward in move 12.
Move 14: **P** chooses a_1

Move 15: **O** challenges the pronoun *he*.
Move 16: **P** can answer a_1*: man*, since **O** conceded it before (namely in move 9)**.**
Move 17: **O** forces **P** to bring forward the identity underlying the pronoun *he*.
Move 18: **P** brings forward the required identity**.**
Move 19: **O** forces **P** to use the identity brought forward in move 18 and apply it to drop the pronoun occurring in 14.
Move 20: **P** drops the pronoun and this yields $R^{\forall}(z)$ *: happy*(a_1).
Mover 21: O asks to resolve the instruction occurring in the last move. Since it is an elementary expression and **O** did not concede it before **P** cannot has no move to play and loses the play.

P has a winning strategy if in a given context, a_1 is a man who smiles and is happy and that stands for every choice of man that **O** can make. Of course, this sentence is not valid. We could develop a material dialogue, by introducing concessions by **O** (premises) and thus check if there is or not a winning strategy. If there is, it amounts to an inference from materially given premises. What we should not do is to design the material dialogue with help of a model (like Hintikka does) this would work against the epistemic frame underlying the present approach.

Let us now come to the analysis of the notorious example of the donkey sentence. We follow the analysis of Sundholm (1986) that constitutes a landmark in the application of CTT to natural language. In order to keep the focus in the interdependence of choices we skip the pronouns *he* and *it* and we replace them with the corresponding instructions already.

Every man who owns a donkey beats it.

As in the example above, first we formalize the first part of the sentence "man who owns a donkey" and we consider that sentence in the context, so we obtain

$$z : (\exists x : man)\ \ (\exists y : donkey)\,(x\ owns\ y)\,.$$

Since the existential *(*$\exists$*y: donkey)(x owns y)* is in fact a compound of the set z, it is more convenient to use the set-separation notation

$$z : \left\{x : man \middle|\ \ (\exists y : donkey)\,(x\ owns\ y)\right\}.$$

We take the left part of z to pick up a man (that owns a donkey). The right part of z is the owned donkey (that is beaten). Putting all together yields

$$p : \left(\forall z : \left\{x : man \middle|\ \ (\exists y : donkey)\,(x\ owns\ y)\right\}\right)\ \left(L^{\{\dots\}}(z) beats\ L^{\exists}\left(R^{\{\dots\}}(z)\right)\right)$$

Or more briefly

$$p : \left(\forall z : \left\{x : M \middle|\ (\exists y : D)\ O\,(x, y)\right\}\right) B\left(L^{\{\dots\}}(z),\ L^{\exists}\left(R^{\{\dots\}}(z)\right)\right),$$

which is nothing else than Sundholm's formalization where instructions take the place of selectors.

Let us run the play but this time with a material dialogue – also here we ignore repetition ranks. Since it is a material dialogue, we know by the formation plays how the sets are composed and we also know that *m* is man, *d* a donkey and that *p'* is a play-object for the proposition that *m owns d.* The point of the thesis is that **P** claims that if we know what has been already mentioned and given *that very man who owns a donkey beats it*, then *man m beats donkey d*.

	O			P	
I	*! M : set*				
II	*! D : set*				
III	*! O(x, y) : set (x : M, y : D)*				
IV	*! Bxy : set (x : M, y : D)*				
V	*! p : (∀z: {x : M | (∃y: D) O(x, y)})B* $(L^{\{\cdots\}}(z), L^{\exists}(R^{\{\cdots\}}(z)))$				
VI	*! m : M*				
VII	*! d : D*				
VIII	*! p' : O(m, d)*				
				! B(m,d)	0
1	*n :=* ...			*m :=* ...	2
3	*?play-object*	(0)		*! q : B(m,d)*	30
25	*!* $R^{\forall}(p)$ *:* $B(L^{\{\cdots\}}(z), L^{\exists}(R^{\{\cdots\}}(z)))$		(V)	*!* $L^{\forall}(p)$ *: {x : M | (∃y: D) O(x, y)}*	4
5	$L^{\forall}(p)/?$	(4)		*! z : {x : M | (∃y: D) O(x, y)}*	6
7	$?_L$	(6)		*!* $L^{\{\cdots\}}(z)$ *: M*	8
9	$L^{\{\cdots\}}(z)/?$	(8)		*! m : M*	10
11	$?_R$	(6)		*!* $R^{\{\cdots\}}(z)$ *: (∃y: D) O(m, y)*	12
13	$R^{\{\cdots\}}(z)/?$	(12)		*!* $(L^{\exists}(R^{\{\cdots\}}(z)), R^{\exists}(R^{\{\cdots\}}(z)))$ *: (∃y: D) O(m, y)*	14
15	$L^{\exists}(R^{\{\cdots\}}(z))/?, R^{\exists}(R^{\{\cdots\}}(z))/?$	(14)		*! (d,p') : (∃y: D) O(m, y)*	16
17	$?_L$	(16)		*!* $L^{\exists}(d,p')$ *: D*	18
19	$L^{\exists}(R^{\{\cdots\}}(z))/?$	(18)		*! d : D*	20
21	$?_R$	(16)		*!* $R^{\exists}(d, p')$ *: O(m, d)*	22
23	$^{R\exists}(R^{\{\cdots\}}(z))/?$	(22)		*! p' : O(m, d)*	24
27	*!* $R^{\forall}(p)$ *:* $B(m,d)$		(25)	$L^{\{\cdots\}}(z)/m, L^{\exists}(R^{\{\cdots\}}(z))/d$	26
29	*! q :* $B(m,d)$		(27)	$R^{\forall}(p)/?$	28

Description

Moves I – VIII: These moves are **O**'s initial concessions. Moves I- IV deal with the formation of expressions. After that the Opponent concedes the donkey sentence and atomic expressions related to the sets *M*, *D* and *O(x, y)*.

Moves 0–3: The Proponent posits the thesis. The players choose their repetition ranks in moves 1 and 2. The actual value they choose does not really matter for

the point we want to illustrate here, thus we simply assume that they are enough for this play and leave them unspecified. Now, when **P** posited the thesis he did not specified the play- objects so **O** asks for it in move 3.

Move 4: **P** chooses to launch a counterattack by challenging the donkey sentence which **O** conceded at V. The rules allow the Proponent to answer directly to the Opponent's first challenge, but then he would not be able to win.

Move 5–24: The dialogue then proceeds in a straightforward way with respect to the rules introduced in Sects. 3.2.2 and 3.2.3. More precisely, this dialogue displays the case where **O** chooses to challenge **P**'s posits as much as she can before answering **P**'s challenge 4.

Notice that the Opponent cannot challenge the Proponent's atomic expressions posited at moves 10, 20 and 24: since **O** made the same posits in her initial concessions VI–VIII, the modified formal rule SR3 forbids her to challenge them.

Move 25: When there is nothing left for her to challenge, **O** comes back to the last unanswered challenge by **P** which was move 4 and makes the relevant defence according to the particle rule for universal quantification.

Moves 26–27: The resolution for instructions $L^{\{\cdots\}}(z)$ *and* $L^{\exists}(R^{\{\cdots\}}(z))$ has been carried out during the dialogue with moves 9–10 and 23–24. Thus the Proponent can use the established substitutions to challenge move 25 according to the structural rule SR4.2. The Opponent defends by performing the requested substitutions.

Moves 28–30: The Proponent then asks the play-object for which the instruction $R^{\forall}(z)$ stands. When she answers, the Opponent posits exactly what **P** needs to defend against **O**'s challenge 3. Notice that at this point this is the last unanswered challenge by **O**, therefore **P** is allowed to answer it in accordance to the structural rule SR1*i*. He does so with his move 30.

Since **O** made the same posit, the rule SR3 forbids her to challenge it. She then has no further possible move, and the Proponent wins this dialogue.

What about the more difficult examples involving *branching quantifiers*?. As shown by Sundholm (2016, forthcoming) a CTT-analysis yields[55]

$$(\exists f \in (\Pi x \in D)\, D)\, (\exists g \in (\Pi x \in D)\, D)\, (\forall x \in D)\, (\forall u \in D)\, A\, [ap\,(f, x)\,/y,\ ap\,(g, u)\,/v]\,.$$

The dialogical development is straightforward if we recall from 3.2.2 that the application of a function follows the following rule

Posit	Challenge	Defence
X! $p : \varphi\ [f(k_1)]$	**Y** $f(k_1)/?$	**X**$p : \varphi\ [k_2/f(k_1]$
		$<\varphi\ [f(k_1)] = \varphi\ [k_2/f(k_1)] : set>$

[55] Sundholm uses the membership sign instead of the colon. For a discussion on this aspect of the notation in CTT, see Granström (2011).

Notice how close this formulation is to the one proposed by Sandu and Jacot (2012): the crucial difference is that in Sundholm's formulation the relevant functions are explicit proof-objects.

3.4.5 Final Remarks

As already mentioned, Hintikka's remark on *binding scope* touches a crucial point in the semantics of anaphora, namely that of structure of dependences, which naturally leads to game-theoretical interpretation. However, on our view, this point involves the dependences between play-objects in general and choices for the substitution of the instructions in particular. A framework such as the one of CTT was necessary to make the point of these dependences explicit.

The dialogical approach implements these dependences within a game-theoretical analysis. Indeed without such an approach choice dependences cannot be expressed at the object language level of first-order logic. As already mentioned the point can be put in the following way: if binding scope amounts to dependences, dependences are understood as interactions, specific forms of the latter represent winning-strategies and we would like to express this in the object language, then it looks natural to embed such structures in a frame where proof-objects are expressed at the object language level as truth-makers of first-order expressions. Moreover, if we go a step deeper in the analysis, it looks natural to introduce a more fundamental semantic level on the basis of which strategies are constructed: this is precisely what the dialogical approach provides by furnishing both play-objects and strategic-objects (the latter are those play-objects relevant for a winning strategy).

More generally, the steps of substitution and identity involving commitments and entitlements so central to Brandom's reading of anaphora cannot be made explicit without such a frame either. Notice that Brandom's (1994, p. 493) analysis of, for example, the donkey sentence, leaves play-object dependences implicit. More precisely, because Brandom does not use an explicit inferential frame such as the one of CTT, he does not distinguish the types *set*, such as *man* and *donkey* over which the subset of *all those men who own a men* is defined (recall that, according to the CTT-reading, *all those men who own a donkey* is an existential defined over the sets *man* and *donkey*).

One other advantage of the dialogical approach is that the meaning of the anaphoric expression is obtained at the play-level and not through the existence of winning strategy for a player as in GTS. That is an advantage because it shows how we can understand the meaning of anaphoric pronoun without knowing how to win the game: it is enough that one understands all the steps the Proponent is committed to in a dialogical game. This is linked to our discussion in the introduction on how the distinction between the play and the strategy seems to provide a way to give shape to Brandom's (1994, p. 636) claim that the "grasp of concepts" amounts to the mastery of inferential roles but this only requires *enough* knowledge of the moves

in the relevant games. The dialogician responds: *enough* means to know the relevant moves that have to be brought forward (according to the local rules) during a play.

3.5 Prospectives

The development of a dialogical approach to CTT is still at its beginnings and many open issues have yet to be tackled. Let us briefly mention only two main research paths that are works in progress:

1. **The Meaning of Conversations:** In his book *The Interactive Stance*, J. Ginzburg (2012) stresses the utmost importance of taking conversational (interactive) aspects into account in order to develop a theory of meaning, where meaning is constituted during the interaction. In order to implement such a theory of meaning Ginzburg makes use of Constructive Type Theory where the so-called "metalogical" rules that constitute meaning are explicitly imported into the object language. Moreover, Ginzburg designs some kind of language games called *dialogical-gameboards* in order to capture the dynamic aspects of every-day dialogues. Now, if we take seriously the claim that meaning is constituted by and within interaction then we expect that the semantics of the underlying logical elements is also understood dialogically. In this context, a dialogical approach to Constructive Type Theory provides both a dialogical frame for the underlying logic and a natural link to the dialogical-gameboards. Rahman (2014) has started tackling this issue but a full development is still to be worked out.
2. **Modal Epistemic Logic and Belief Revision:** In the context of CTT, the variable in a hypothetical such as $p(x) : P\ (x : S)$ represents *an unknown element of S* that can be instantiated by some s when the required knowledge is available.[56] Thus, in this framework, instantiating the *unknown* element x by some s *known* to be a fixed (but arbitrary) element of S describes the passage from belief to knowledge. Using the current terminology of epistemic logic as an analogy – in the style of Hintikka (1962) – we say that a judgement of the form $x : S$ expresses *belief* rather than *knowledge*. In fact, for this transition to count as a transition to knowledge, it is not only necessary that $s : S$, but it is also necessary that the proof-object s is of the adequate sort.[57] In other words, we also need to have the definition $x{=}s : S$. This definition of x can be called an *anchoring* of the hypothesis (belief) S in the *actual* world.[58] Thus, the result of this anchoring process yields $p(x{=}s) : P\ (s : S)$. In fact after some seminal work of Aarne Ranta (1991) there are ongoing developments by Giuseppe Primiero

[56]Cf. Granström (2011, pp. 110–112). In fact, chapter V of Granström (2011) contains a thorough discussion of the issue.

[57]Cf. Ranta (1994, pp. 151–154).

[58]Cf. Ranta (1994, p. 152).

(2008, 2012) on applying CTT to belief revision. However neither have the dynamic aspects provided by game-theoretical approaches been considered – where knowledge acquisition is depicted as resulting from interaction – nor the modal formalizations of belief revision have been yet studied in this framework. B. Dango has started to work out the ways to combine the CTT formulation of modal logic with the dialogical approach.

Appendix: Standard Dialogical Games

Let L be a first-order language built as usual upon the propositional connectives, the quantifiers, a denumerable set of individual variables, a denumerable set of individual constants and a denumerable set of predicate symbols (each with a fixed arity).

We extend the language L with two labels ***O*** and ***P***, standing for the players of the game, and the two symbols '*!*' and '?'. When the identity of the player does not matter, we use variables ***X*** or ***Y*** (with $X \neq Y$). A move is an expression of the form '***X***-*e*', where *e* is either of the form '$!\varphi$' for some sentence φ of L or of the form '$?[!\varphi_1, \ldots, !\varphi_n]$'.

The *particle* (or local) *rules* for standard dialogical games are given in the following table:

Previous move	$\mathbf{X}!\varphi \wedge \psi$	$\mathbf{X}!\varphi \vee \psi$	$\mathbf{X}!\varphi \rightarrow \psi$	$\mathbf{X}! \neg \varphi$
Challenge	$\mathbf{Y}?[!\varphi]$ or $\mathbf{Y}?[!\psi]$	$\mathbf{Y}?[!\varphi, !\psi]$	$\mathbf{Y}!\varphi$	$\mathbf{Y}!\varphi$
Defence	$\mathbf{X}!\varphi$ resp. $\mathbf{X}!\psi$	$\mathbf{X}!\varphi$ or $\mathbf{X}!\psi$	$\mathbf{X}!\psi$	– –

Previous move	$\mathbf{X}!\forall x\varphi$	$\mathbf{X}!\exists x\varphi$
Challenge	$\mathbf{Y}?[!\varphi(x/a_i)]$	$\mathbf{Y}$-$[!\varphi(x/a_1), \ldots, !\varphi(x/a_n)]$
Defence	$\mathbf{X}!\varphi(x/a_i)$	$\mathbf{X}!\varphi(x/a_i)$ with $1 \leq i \leq n$

In this table, the a_is are individual constants and $\varphi(x/a_i)$ denotes the formula obtained by replacing every free occurrence of x in φ by a_i. When a move consists in a question of the form '$?[!\varphi_1, \ldots, !\varphi_n]$', the other player chooses one formula among $\varphi_1, \ldots, \varphi_n$ and plays it. We thus distinguish conjunction from disjunction and universal quantification from existential quantification in terms of which player chooses. With conjunction and universal quantification, the challenger chooses

which formula he asks for. With disjunction and existential quantification, it is the defender who can choose between various formulas. Notice that there is no defence in the particle rule for negation.

Particle rules provide an abstract description of how the game can proceed locally: they specify the way a formula can be challenged and defended according to its main logical constant. In this way the particle rules govern the local level of meaning. Strictly speaking, the expressions occurring in the table above are not actual moves because they feature formula schemata and the players are not specified. Moreover, these rules are indifferent to any particular situations that might occur during the game. For these reasons we say that the description provided by the particle rules is abstract.

Since the players' identities are not specified in these rules, particle rules are symmetric: the rules are the same for the two players. The local meaning being symmetric (in this sense) is one of the greatest strengths of the dialogical approach to meaning. It is in particular the reason why the dialogical approach is immune to a wide range of trivializing connectives such as Prior's *tonk*.[59]

The expressions occurring in particle rules are all move schematas. The words "challenge" and "defence" are convenient to name certain moves according to their relation with other moves which can be defined in the following way. Let σ be a sequence of moves. The function p_σ assigns a position to each move in σ, starting with 0. The function F_σ assigns a pair $[m,Z]$ to certain moves N in σ, where m denotes a position smaller than $p_\sigma(N)$ and Z is either C or D, standing respectively for "challenge" and "defence". That is, the function F_σ keeps track of the relations of challenge and defence as they are given by the particle rules. Consider for example the following sequence σ:

$$\mathbf{P}!\varphi \wedge \psi,\ \mathbf{P}!\chi \wedge \psi,\ \mathbf{O}?\,[!\varphi]\,,\ \mathbf{P}!\varphi$$

In this sequence we have for example $p_\sigma(\mathbf{P}!\chi\wedge\psi)=1$.

A *play* is a legal sequence of moves, i.e., a sequence of moves which observes the game rules. Particle rules are not the only rules which must be observed in this respect. In fact, it can be said that the second kind of rules named *structural rules* are the ones giving the precise conditions under which a given sequence is a play. The dialogical game for φ, written $\mathsf{D}(\varphi)$, is the set of all plays with φ being the *thesis* (see the Starting rule below). The structural rules are the following:

SR0 (Starting rule). Let φ be a complex sentence of L and i,j be positive integers. For every $\zeta\in\mathsf{D}(\varphi)$ we have:

- $p_\zeta(\mathbf{P}!\,\varphi)=0$,
- $p_\zeta(\mathbf{O}\ \mathrm{n}:=i)=1$,
- $p_\zeta(\mathbf{P}\ \mathrm{m}:=j)=2$.

[59] See Rahman et al. (2009) and Rahman (2012).

In other words, any play ζ in $\mathsf{D}(\varphi)$ starts with **P** positing φ. We call φ the thesis of both the play and the dialogical game. After that, the Opponent and the Proponent successively choose a positive integer called repetition rank. The role of these integers is to ensure that every play ends after finitely many moves in the way specified by the next structural rule.

SR1 (Classical game-playing rule)

- Let $\zeta \in \mathsf{D}(\varphi)$. For every Min ζ with $p_\zeta(M)>2$ we have $F_\zeta(M)=[m',Z]$ with $m'<p_\zeta(M)$ and $Z \in \{C,D\}$.
- Let r be the repetition rank of player ***X*** and $\zeta \in \mathsf{D}(\varphi)$ such that
  - the last member of ζ is a ***Y***-move,
  - M_0 is a ***Y***-move of position m_0 in ζ,
  - M1, . . . ,Mn are ***X***-moves in ζ such that $F_\zeta(M_1)= \ldots F_\zeta(M_n)=[m_0,Z]$.

Consider the sequence[60] $\zeta'=\zeta^*N$ where N is an ***X***-move such that $F_{\zeta'}(N)=[m_0,Z]$. We have $\zeta' \in \mathsf{D}(\varphi)$ only if n<r.

The first part of the rule states that every move after the repetition rank choices is either a challenge or a defence. The second part ensures finiteness of plays by setting the player's repetition rank as the maximum number of times he can challenge or defend against a given move by the other player.

SR2 (Formal rule)
Let ψ be an elementary sentence, N be the move $\mathbf{P}!\psi$ and M be the move $\mathbf{O}!\psi$. A sequence ζ of moves is a play only if we have: if $N \in \zeta$ then $M \in \zeta$ and $p_\zeta(M)<p_\zeta(N)$.

That is, the Proponent can play an elementary sentence only if the Opponent has played it previously. The Formal rule is one of the characteristic features of the dialogical approach: other game-based approaches do not have it.

Helge Rückert pointed out that the formal rule triggers a novel notion of validity: *Geltung* (Legitimacy).[61] Indeed with this rule the dialogical framework comes with an internal account for elementary sentences: an account in terms of interaction only, without depending on metalogical meaning explanations for the non-logical vocabulary. More prominently this means that the dialogical account does not rely – contrary to Hintikka's GTS games – on the model-theoretical approach to meaning for atomic formulas.

From there Rückert claims, and on this point we disagree with him, that Geltung is the idea that interaction emerges without knowing (or without needing to know) what the meaning of elementary sentences are. We disagree because the question of the meaning of elementary sentences (and more generally, of non-logical vocabulary) cannot be disregarded if the dialogical framework is meant to provide a general theory of meaning. In our view, thus, Rückert's interpretation of Geltung dissolves the meaning of elementary sentences into the formal rule. This is mainly

[60] We use ζ^*N to denote the sequence obtained by adding move N to the play ζ.

[61] Rückert (2011b).

due to the fact that the standard version of the framework does not have the means to express a semantic at the object-language level in terms of asking and giving reasons for elementary sentences. As a consequence, the standard formulation simply relies on the formal rule which amounts to entitle **P** to copy-cat the elementary sentences brought forward by **O**. According to us, the introduction of play-objects provides a solution to this without giving up on the internal aspect linked with Geltung. We will develop this idea when we give the particle rules in Sect. 3.2.3 and after we introduce a "modified formal rule" in Sect. 3.2.4.

Here is some terminology for the last structural rule in standard dialogical games. A play is called *terminal* when it cannot be extended by further moves in compliance with the rules. We say it is **X**-terminal when the last move in the play is an **X**-move.

SR3 (Winning Rule). Player **X** wins the play ζ only if it is **X**-terminal.

Consider for example the following sequences of moves:

$$\mathbf{P}-Qa\wedge Qb,\ \mathbf{O}-\mathrm{n}:=1,\ \mathbf{P}-\mathrm{m}:=6,\ \mathbf{O}-?\,[Qa]\,,\ \mathbf{P}-Qa$$
$$\mathbf{P}-Qa\rightarrow Qa,\ \mathbf{O}-\mathrm{n}:=1,\ \mathbf{P}-\mathrm{m}:=12,\ \mathbf{O}-\mathrm{Qa},\ \mathbf{P}-Qa$$

The first one is not a play because it breaks the Formal rule: with his last move, the Proponent plays an elementary sentence which the Opponent has not played beforehand. By contrast, the second sequence is a play in D(**P**-$Qa\rightarrow Qa$). We often use a convenient table notation for plays. For example, we can write this play as follows:

	O		P	
			! Qa→Qa	0
1	n:=1		m:=12	2
3	*! Qa*	(0)	*! Qa*	4

The numbers in the external columns are the positions of the moves in the play. When a move is a challenge, the position of the challenged move is indicated in the internal columns, as with move 3 in this example. Notice that such tables carry the information given by the functions p and F in addition to represent the play itself.

However, when we want to consider several plays together – for example when building a strategy – such tables are not that perspicuous. So we do not use them to deal with dialogical games for which we prefer another perspective. The *extensive form* of the dialogical game D(φ) is simply the tree representation of it, also often called the game-tree. More precisely, the extensive form E_φ of D(φ) is the tree (T,l,S) such that:

(i) Every node t in T is labelled with a move occurring in $\mathsf{D}(\varphi)$
(ii) l: T→**N**
(iii) S⊆T^2 with:

- There is a unique t_0 (the root) in T such that $l(t_0)=0$, and t_0 is labelled with the thesis of the game,
- For every $t\neq t_0$ there is a unique t' such that t'St,
- For every t and t' in T, if tSt' then $l(t')=l(t)+1$,
- Let $\zeta\in\mathsf{D}(\varphi)$ such that $p_\zeta(M')=p_\zeta(M)+1$. If t and t' are respectively labelled with M and M', then tSt'.

Many dialogical game metalogical results are obtained by leaving the level of rules and plays to move to the level of strategies. Significant among these results are the ones concerning the existence of winning strategies for a player. We will now define these notions and give examples of such results.

A *strategy* for player **X** in $\mathsf{D}(\varphi)$ is a function which assigns an **X**-move M to every non terminal play ζ having a **Y**-move as last member such that extending ζ with M results in a play. An **X**-strategy is *winning* if playing according to it leads to **X**'s victory no matter how **Y** plays.

Strategies can be considered from the perspective of extensive forms: the extensive form of an **X**-strategy s in $\mathsf{D}(\varphi)$ is the tree-fragment $S_\varphi=(T_s,l_s,S_s)$ of E_φ such that:

(i) The root of S_φ is the root of E_φ,
(ii) Given a node t in E_φ labelled with an **X**-move, we have $t'\in T_s$ and tS_st' whenever tSt'.
(iii) Given a node t in E_φ labelled with a **Y**-move and with at least one t' such that tSt', we have a unique s(t) in T_s with $tS_ss(t)$ and s(t) is labelled with the **X**-move prescribed by s.

Here are some results pertaining to the level of strategies[62]:

- Winning **P**-strategies and leaves. *Let* w *be a winning* **P***-strategy in* $\mathsf{D}(\varphi)$*. Then every leaf in the extensive form* W_φ *of* w *is labelled with a* **P** *elementary sentence.*
- Determinacy. *There is a winning* **X***-strategy in* $\mathsf{D}(\varphi)$ *if and only if there is no winning* **Y***-strategy in* $\mathsf{D}(\varphi)$*.*
- Soundness and Completeness of Tableaux. *Consider first-order tableaux and first-order dialogical games. There is a tableau proof for* φ *if and only if there is a winning* **P***-strategy in* $\mathsf{D}(\varphi)$*.*

The fact that existence of a winning **P**-strategy coincides with validity (*there is a winning* **P***-strategy in* $\mathsf{D}(\varphi)$ *if and only if* φ *is valid)* follows from the soundness and completeness of the tableau method with respect to model-theoretical semantics.

Regarding several results, extensive forms of strategies have key parts: one of the parts of a winning strategy, called the *core* of the strategy, is actually that on

[62]These results are proven, together with others, in Clerbout (2014a).

which one works when considering translation algorithms such as the procedures. The basic idea behind the notion of core is to get rid of redundant information (for example, different orders of moves) which we find in extensive forms of strategies (see Clerbout and Rahman (2015).

References

Abramsky, S., Mellies, P.A.: Concurrent games and full completeness. In: Proceedings of the 14th International Symposium on Logic in Computer Science, pp. 431–442, IEEE Computer Society Press, Trento (1999)

Austin, J.L.: Other minds. Proc. Aristot. Soc., Supplementary Volume **20**, 148–187 (1946)

Bell, J.: The Axiom of Choice. College Publications, London (2009)

Blass, A.: A game semantics for linear logic. Ann. Pure. Appl. Logic. **56**, 183–220 (1992)

Brandom, R.: Making It Explicit. Harvard University Press, Cambridge, MA/London (1994)

Brandom, R.: Articulating Reasons. Harvard University Press, Cambridge, MA/London (2000)

Clerbout, N.: First-order dialogical games and tableaux. J. Philos. Logic. **43**(4), 785–801. doi:10.1007/s10992-013-9289-z. URL http://dx.doi.org/10.1007/s10992-013-9289-z (2014a)

Clerbout, N.: La sémantique dialogique: Concepts fondamentaux et éléments de metathéorie. College Publications, London (2014b)

Clerbout, N.: Finiteness of plays and the dialogical problem of decidability. IfCoLog J. Logics. Appl. **1**(1), 115–130 (2014c)

Clerbout, N., Rahman, S.: On dialogues, predication and elementary sentences. Revista de Humanidades de Valparaíso **2**, 7–46 (2013)

Clerbout, N., Rahman, S.: Linking Game-Theoretical Approaches with Constructive Type Theory: Dialogical Strategies, CTT Demonstrations and the Axiom of Choice. Springer, Dordrecht, in print (2015)

Clerbout, N., Gorisse, M.H., Rahman, S.: Context-sensitivity in Jain philosophy: a dialogical study of Siddharsigani's commentary on the handbook of logic. J. Philos. Log. **40**(5), 633–662 (2011)

Enderton, H.: Finite partially ordered quantifiers. Zeitschrift für mathematische Logik **16**, 393–397 (1970)

Feferman, S.: What kind of logic is "Independence Friendly" logic?. In: Auxier, R.E., Hahn, L.E. (eds.) The Philosophy of Jaakko Hintikka, vol. 30, pp. 453–469. Open Court – Library of Living Philosophers, Chicago (2006)

Felscher, W.: Dialogues as a foundation for intuitionistic logic. In: Gabbay, D., Guenthner, F. (eds.) Handbook of Philosophical Logic, vol. 3, pp. 341–372. Kluwer, Dordrecht (1985)

Felscher, W.: Review of Jean E. Rubin 'Mathematical logic: applications and theory'. J. Symb. Log. **59**, 670–671 (1994)

Fiutek, V., Rückert, H., Rahman, S.: A dialogical semantics for Bonanno's system of belief revision. In: Bour, P., alii (eds.) Constructions, pp. 315–334. College Publications, London (2010)

Fontaine, M.: Argumentation et engagement ontologique. Être, c'est être choisi. College Publications, London (2013)

Ginzburg, J.: The Interactive Stance: Meaning for Conversation. Oxford University Press, Oxford (2012)

Girard, J.Y.: On the meaning of logical rules I: syntax vs. semantics. In: Berger, U., Schwichtenberg, H. (eds.) Computational Logic, pp. 215–272. Springer, Heidelberg (1999)

Granström, J.: Treatise on Intuitionistic Type Theory. Springer, Dordrecht (2011)

Hintikka, J.: Knowledge and Belief. Cornell University Press, Ithaca (1962)

Hintikka, J.: Logic, Language-Games and Information: Kantian Themes in the Philosophy of Logic. Clarendon, Oxford (1973)
Hintikka, J.: The Principles of Mathematics Revisited. Cambridge University Press, Cambridge (1996a)
Hintikka, J.: Lingua Universalis vs. Calculus Ratiocinator: An Ultimate Presupposition of Twentieth-Century Philosophy. Kluwer, Dordrecht (1996b)
Hintikka, J.: Defining Truth, the Whole Truth and Nothing but the Truth, Reports from the Department of Philosophy, no. 2. University of Helsinki, Helsinki (1997)
Hintikka, J.: Inquiry as Inquiry: A Logic of Scientific Discovery. Springer, Dordrecht (1999)
Hintikka, J., Kulas, J.: Anaphora and Definite Descriptions, Two Applications of Game-Theoretical Semantics. Reidel, Dordrecht/Boston/Lancaster (1985)
Hintikka, J., Sandu, G.: Informational independence as a semantical phenomenon. In: Fenstad, J.E., Frolov, I.T., Hilpinen, R. (eds.) Logic, Methodology and Philosophy of Science, vol. 8, pp. 571–589. Elsevier, Amsterdam (1989)
Hintikka, J., Sandu, G.: Game-theoretical semantics. In: van Benthem, J., ter Meulen, A. (eds.) Handbook of Logic and Language, pp. 361–410. Elsevier, Amsterdam (1997)
Hintikka, J., Halonen, I., Mutanen, A.: Interrogative logic as a general theory of reasoning. In Hintikka (1999), pp. 47–90 (1999)
Hodges, W.: Dialogue foundations. a sceptical look. Proc. Aristot. Soc., supplementary volume **LXXV**, 17–32 (2001)
Hodges, W: Logic and games. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy. URL http://plato.stanford.edu/archives/spr2013/entries/logic-games/ (2004, revised 2013)
Jovanovic, R.: Hintikka's take on the axiom of choice and the constructivist challenge. Revista de Humanidades de Valparaíso **2**, 135–152 (2013)
Jovanovic, R.: Hintikka's Take on the Axiom of Choice and the Constructivist Challenge. College Publications, London (2015)
Kamlah, W., Lorenzen, P.: Logische Propädeutik, 2nd edn. Metzler, Stuttgart/Weimar (1972)
Kamlah, W., Lorenzen, P.: Logical Propaedeutic. University Press of America, Lanham. English translation of Kamlah/Lorenzen (1972) by H. Robinson (1984)
Keiff, L.: Heuristique formelle et logiques modales non-normales. Philos. Sci. **8**(2), 39–57 (2004a)
Keiff, L.: Introduction à la logiaue modale et hybride. Philos. Sci. **8**(2), 89–102 (2004b)
Keiff, L.: Le Pluralisme Dialogique: Approches dynamiques de l'argumentation formelle. Ph.D. thesis, Lille 3, Lille (2007)
Keiff, L.: Dialogical logic. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy. URL http://plato.stanford.edu/entries/logic-dialogical/ (2009)
Lecomte, A., Quatrini, M.: Pour une étude du langage via l'interaction: dialogues et sémantique en Ludique. Mathématiques et Sciences Humaines **189**, 37–67 (2010)
Lecomte, A.: Meaning, Logic and Ludics. Imperial College Press, London (2011)
Lecomte, A., Quatrini, M.: Figures of dialogue: a view from ludics. Synthese **183**(1), 59–85 (2011)
Lecomte, A., Tronçon, S. (eds.): Ludics, Dialogues and Interaction: PRELUDE Project 2006–2009. Revised Selected Papers. Springer, Berlin/Heidelberg (2011)
Lorenz, K.: Elemente der Sprachkritik. Eine Alternative zum Dogmatismus und Skeptizismus in der Analytischen Philosophie. Suhrkamp, Frankfurt (1970)
Lorenz, K.: Basic objectives of dialogue logic in historical perspective. In: Rahman, S., Rückert, H. (eds.) New Perspectives in Dialogical Logic, special volume Synthese **127**(1–2), 255–263 (2001)
Lorenz, K.: Dialogischer Konstruktivismus. De Gruyter, Berlin (2008)
Lorenz, K.: Logic, Language and Method: On Polarities in Human Experience. De Gruyter, Berlin/New York (2010a)
Lorenz, K.: Philosophische Variationen: Gesammelte Aufstze unter Einschluss gemeinsam mit Jurgen Mittelstraß geschriebener Arbeiten zu Platon und Leibniz. De Gruyter, Berlin/New York (2010b)
Lorenzen, P.: Einführung in die operative Logik und Mathematik. Springer, Berlin (1955)

Lorenzen, P., Lorenz, K.: Dialogische Logik. Wissenschaftliche Buchgesellschaft, Darmstadt (1978)
Lorenzen, P., Schwemmer, O.: Konstruktive Logik, Ethik und Wissenschaftstheorie, second edition. Bibliographisches Institut, Mannheim (1975)
Magnier, S.: Approche dialogique de la dynamique épistémique et de la condition juridique. College Publications, London (2013)
Marion, M.: Hintikka on Wittgenstein: from language-games to game semantics. In: Aho, T., Pietarinen, A.-V. (eds.) Truth and Games, Essays in Honour of Gabriel Sandu, Acta Philosophica Fennica, vol. 78, pp. 223–242. Helsinki : Societas Philosophica Fennica (2006)
Marion, M.: Why play logical games? In: Majer, O., Pietarinen, A.-V., Tulenheimo, T. (eds.) Logic and Games, Foundational Perspectives, pp. 3–26. Springer, Dordrecht (2009)
Marion, M.: Between saying and doing: from Lorenzen to Brandom and Back. In: Bour, P.E., Rebuschi, M., Rollet, L. (eds.) Constructions, pp. 489–497. College Publications, London (2010)
Martin-Löf, P.: Intuitionistic Type Theory. Notes by Giovanni Sambin of a series of lectures given in Padua, June 1980. Bibliopolis, Naples (1984)
Martin-Löf, P.: 100 years of Zermelo's axiom of choice: what was the problem with it? Comput. J. **49**(3), 345–350 (2006)
Penco, C.: Keeping track of individuals. Brandom's analysis of Kripke's puzzle and the content of belief". Pragmat. Cogn. **13**(1), 177–201 (2005)
Popek, A.: Logical dialogues from middle ages. In: Barés Gómez, C., Magnier, S., Salguero, F.J. (eds.) Logic of Knowledge. Theory and Applications, pp. 223–244. College Publications, London (2012)
Prawitz, D.: Proofs and the meaning and completeness of the logical constants. In: Hintikka, J., Niiniluoto, I., Saarinen, E. (eds.) Essays on Mathematical and Philosophical Logic, pp. 25–40. Reidel, Dordrecht (1979)
Prawitz, D.: Truth and proof in intuitionism. In: Dybjer, P., Lindström, S., Palmgren, E., Sundholm, G. (eds.) Epistemology Versus Ontology: Essays on the Philosophy and Foundations of Mathematics in Honour of Per Martin-Löf, pp. 45–68. Springer, Dordrecht (2012)
Primiero, G.: Constructive modalities for information. Talk given at the Young Researchers Days in Logic, Philosophy and History of Science, Brussels, 1–2 Sept 2008 (2008)
Primiero, G.: A contextual type theory with judgemental modalities for reasoning from open assumptions. Log. Anal. **220**, 579–600 (2012)
Rahman, S.: Über Dialogue, Protologische Kategorien und andere Seltenheiten. P. Lang, Frankfurt/Paris/New York (1993)
Rahman, S.: Non-normal dialogics for a wonderful world and more. In: van Benthem, J., alii (eds.) The Age of Alternative Logics: Assessing Philosophy of Logic and Mathematics Today, pp. 311–334. Springer, Dordrecht (2009)
Rahman, S.: Negation in the logic of first degree entailment and *tonk*: a dialogical study. In: Rahman, S., Primiero, G., Marion, M. (eds.) The Realism-Antirealism Debate in the Age of Alternative Logics, pp. 213–250. Springer, Dordrecht (2012)
Rahman, S.: From dialogue to dialogue: Conversations and the dialogical approach to meaning. In: Bowao, C., Rahman, S. (ed.) De l'orature à l'écriture, pp. 70–106. King's College/College Publications, London (2014)
Rahman, S.: On hypothetical judgements and Leibniz's notion of conditional right. In: Armgardt, A., Canivez, P., Chassagnard-Pinet, S. (eds.) Legal Reasoning and Logic. Past & Present Interactions, pp. 109–167. Springer, Dordrecht (2015)
Rahman, S., Clerbout, N.: Constructive type theory and the dialogical approach to meaning. The Baltic International Yearbook of Cognition, Logic and Communication: Games, Game Theory and Game Semantics, vol. 8, pp. 1–72. Also online in: www.thebalticyearbook.org (2013)
Rahman, S., Clerbout, N.: Constructive type theory and the dialogical turn – a new start for Erlangen constructivism. In: Mittelstrass, J., von Bülow, C. (eds.) Dialogische Logik, pp. 127–184. Mentis, Münster (2015)

Rahman, S., Keiff, L.: On how to be a dialogician. In: Vanderveken, D. (ed.) Logic, Thought, and Action, pp. 359–408. Kluwer, Dordrecht (2005)
Rahman, S., Keiff, L.: La Dialectique entre logique et rhetorique. Rev. Metaphys. Morale **66**(2), 149–178 (2010)
Rahman, S., Primiero, G., Marion, M. (eds.): The Realism-Antirealism Debate in the Age of Alternative Logics. Springer, Dordrecht (2012)
Rahman, S., Redmond, J.: Armonía Dialógica: *tonk* Teoría Constructive de Tipos y Reglas para Jugadores Anónimos, Constructive type theory and player-independent rules. Theoria, 2015, in print (2015a)
Rahman, S., Redmond, J.: A dialogical frame for fictions as hypothetical objects. UNISINOS, **16**(1), pp. 2–21 (2015b)
Rahman, S., Tulenheimo, T.: From games to dialogues and back: towards a general frame for validity. In: Majer, O., Pietarinen, A., Tulenheimo, T. (eds.) Games: Unifying Logic, Language and Philosophy, pp. 153–208. Springer, Dordrecht (2009)
Rahman, S., Clerbout, N., Keiff, L.: On dialogues and natural deduction. In: Rahman, S., Primiero, G. (eds.) Acts of Knowledge: History, Philosophy and Logic, pp. 301–336. College Publications, London (2009)
Rahman, S., Clerbout, N., McConaughey, Z.: On play-objects in dialogical games. Towards a dialogical approach to constructive type theory. In: Allo, P., Kerkhove, V.v. (ed.) Modestly radical or radically modest. Festschrift for Jean-Paul van Bendegem, pp. 127–154. College Publications, London (2014)
Ranta, A.: Propositions as games as types. Synthese **76**, 377–395 (1988)
Ranta, A.: Constructing possible worlds. Theoria **57**(1–2), 77–99 (1991)
Ranta, A.: Type-Theoretical Grammar. Clarendon, Oxford (1994)
Read, S.: Harmony and modality. In: Dégremont, C., Keiff, L., Rückert, H. (eds.) Dialogues, Logics and Other Strange Things: Essays in Honour of Shahid Rahman, pp. 285–303. College Publications, London (2008)
Read, S.: General elimination harmony and the meaning of the logical constants. J. Philos. Log. **39**(5), 557–576 (2010)
Redmond, J.: Logique dynamique de la fiction: Pour une approche dialogique. College Publications, London (2010)
Redmond, J., Fontaine, M.: How to Play Dialogues: An Introduction to Dialogical Logic. College Publications, London (2011)
Rückert, H.: Dialogues as a Dynamic Framework for Logic. College Publications, London (2011a)
Rückert, H.: The conception of validity in dialogical logic. Talk at the workshop Proofs and Dialogues, Tübingen (2011b)
Sandu, G.: Studies in Game-Theoretical Logics and Semantics. Ph.D. thesis, University of Helsinki, Helsinki (1991)
Sandu, G.: On the theory of anaphora: dynamic predicate logic vs. game-theoretical semantics. Linguist. Philos. **20**, 147–174 (1997)
Sandu, G., Jacot, J.: Quantification and anaphora in natural language. In: Schantz, R. (ed.) Prospect for Meaning, pp. 609–628. Walter de Gruyter Inc, Berlin/New York (2012)
Schröder-Heister, P.: Lorenzen's operative justification of intuitionistic logic. In: Bourdeau, M., van Atten, M., Boldini, P. (eds.) One Hundred Years of Intuitionism (1907–2007), pp. 214–240. Birkhäuser, Basel (2008)
Sterling, J.M.: Note on Diaconescu's theorem. Online in http://www.jonmsterling.com/posts/2015-04-24-note-on-diaconescus-theorem.html (2015)
Sundholm, G.: Proof-theory and meaning. In: Gabbay, D., Guenthner, F. (eds.) Handbook of Philosophical Logic, vol. 3, pp. 471–506. Reidel, Dordrecht (1986)
Sundholm, G.: Implicit epistemic aspects of constructive logic. J. Log. Lang. Inf. **6**(2), 191–212 (1997)
Sundholm, G.: Inference versus consequence. In: Childers, T. (ed.) The Logica Yearbook 1997, pp. 26–36. Filosofia, Prague (1998)

Sundholm, G.: A plea for logical atavism. In: Majer, O. (ed.) The Logica Yearbook 2000, pp. 151–162. Filosofia, Prague (2001)
Sundholm, G.: A century of judgement and inference: 1837–1936. In: Haaparanta, L. (ed.) The Development of Modern Logic, pp. 263–317. Oxford University Press, Oxford (2009)
Sundholm, G.: Independence friendly language is first order after all? Logique et Analyse, in print (2016, forthcoming)
Sundholm, G.: Inference and consequence in an interpeted language. Talk at the Workshop Proof Theory and Philosophy, Groningen, 3–5 Dec, 2013 (2003)
Tait, W.: The law of excluded middle and the axiom of choice. In: George, A. (ed.) Mathematics and Mind, pp. 45–70. Oxford University Press, New York (1994)
Tulenheimo, T.: Independence friendly logic. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy. URL http://plato.stanford.edu/entries/logic-if/ (2009)
Tulenheimo, T.: On some logic games in their philosophical context. In: Lecomte, A., Tronçon, S. (eds.) Ludics, Dialogues and Interaction: PRELUDE Project 2006–2009, Revised Selected Papers, pp. 88–113. Springer, Berlin/Heidelberg (2011)
Väänänen, J.: Second-order logic and foundations of mathematics. Bull. Symbolic. Logic. **7**, 504–520 (2001)
Van Benthem, J.: Exploring Logical Dynamics. CSLI Publications and Cambridge University Press, Stanford/Cambridge (1996)
Van Benthem, J.: Correspondence theory. In: Gabbay, D., Guenthner, F. (eds.) Handbook of Philosophical Logic, vol. 2, pp. 167–247. Reidel, Dordrecht, 1984. Reprint with addenda in second edition, pp. 325–408 (2001)
Van Benthem, J.: Logical Dynamics of Information and Interaction. Cambridge University Press, Cambridge (2011)
Van Benthem, J.: Logic in Games. MIT, Cambridge, MA (2014)
Van Ditmarsch, H., van der Hoek, W., Kooi, B.: Dynamic Epistemic Logic. Springer, Berlin (2007)
Walkoe, W.: Finite partially ordered quantification. J. Symb. Log. **35**, 535–555 (1970)

Chapter 4
Dependent Types for Pragmatics

Darryl McAdams and Jonathan Sterling

Abstract In this paper, we present an extension to Martin-Löf's Intuitionistic Type Theory which gives natural solutions to problems in pragmatics, such as pronominal reference and presupposition. Our approach also gives a simple account of donkey anaphora without resorting to exotic scope extension of the sort used in Discourse Representation Theory and Dynamic Semantics, thanks to the proof-relevant nature of type theory.

Keywords Semantics • Pragmatics • Pronouns • Presuppositions • Type theory • Dependent types • Intuitionism

4.1 Introduction

To begin with, we give a brief overview of the meaning explanations for Intuitionistic Type Theory in Sect. 4.2, and introduce the standard connectives. Section 4.3 establishes the intended meanings of pronouns and determiners under the dependent typing discipline, and introduces an extension to the type theory (namely our *require* rule) which assigns them these meanings in the general case. We first give a computational justification of *require* in light of the meaning explanation, and then give a proof-theoretic justification by showing how to eliminate require expressions from terms by induction on the demonstrations of their well-typedness.

Finally, Sect. 4.4 wraps up with a discussion of further extensions that could be made to the framework, on both theoretical and empirical grounds.

D. McAdams (✉)
2225 NW 6th Terrace, Wilton Manors, FL 33311, USA
e-mail: darryl@languagengine.co

J. Sterling
SlamData, Inc., Boulder, CO, USA
e-mail: jon@jonmsterling.com

J. Redmond et al. (eds.), *Epistemology, Knowledge and the Impact of Interaction*,
Logic, Epistemology, and the Unity of Science 38, DOI 10.1007/978-3-319-26506-3_4

4.2 Type Theory and Its Meaning Explanation

Intuitionistic Type Theory is an approach to first-order and higher-order logic, based on a computational justification called the *verificationist meaning explanation*. First, an untyped and open-ended programming language (also called a computation system) is established with a big-step operational semantics, given by the judgment $M \Rightarrow M'$. Then, a type is defined by specifying how to form a canonical member ("verification"), and when two such canonical members are considered equal. Finally, membership $M \in A$ is evident when $M \Rightarrow M'$ such that M' is a canonical member of A.

In this setting, then, the introduction rules follow directly from the definitions of the types, and the elimination rules are explained by showing how one may transform the evidence for their premises into the evidence for their conclusions. For a more detailed exposition of the verificationist meaning explanation for intuitionistic first order logic, see Martin-Löf (1996); the meaning explanation for full dependent type theory is given in Martin-Löf (1982) and Martin-Löf (1984).

4.2.1 The Connectives of Type Theory

The two main connectives of type theory are the dependent pair $(x : A) \times B$ and the dependent function $(x : A) \to B$, where x may occur free in B.[1]

4.2.1.1 Dependent Pairs

To define the dependent pair type, we first introduce several new terms into the computation system, together with their canonical forms:

$$\frac{}{(x : A) \times B \Rightarrow (x : A) \times B} \qquad \frac{}{\langle M, N\rangle \Rightarrow \langle M, N\rangle}$$

$$\frac{P \Rightarrow \langle M, N\rangle \qquad M \Rightarrow M'}{\mathsf{fst}(P) \Rightarrow M'} \qquad \frac{P \Rightarrow \langle M, N\rangle \qquad N \Rightarrow N'}{\mathsf{snd}(P) \Rightarrow N'}$$

Then, we define the type $(x : A) \times B$ (presupposing A *type* and $x : A \vdash B$ *type*) by declaring $\langle M, N\rangle$ to be a canonical member under the circumstances that $M \in A$ and $N \in [M/x]B$, where $[M/x]B$ stands for the substitution of N for x in B; moreover,

[1] In this paper, we opt to use the notation $(x : A) \times B$ and $(x : A) \to B$ in place of the more common $\Sigma x : A.B$ and $\Pi x : A.B$, respectively, in order to emphasize that these are merely dependent versions of pairs and functions. This notation was first invented in the Nuprl System (Constable et al. 1986).

$\langle M, N\rangle$ and $\langle M', N'\rangle$ are equal canonical members in case $M = M' \in A$ and $N = N' \in [M/x]B$.

The formation and introduction rules for dependent pairs are immediately evident by this definition:

$$\frac{\Gamma \vdash A\ type \qquad \Gamma, x : A \vdash B\ type}{\Gamma \vdash (x : A) \times B\ type}\times\mathrm{F} \qquad \frac{\Gamma \vdash M \in A \qquad \Gamma \vdash N \in [M/x]B}{\Gamma \vdash \langle M, N\rangle \in (x : A) \times B}\times\mathrm{I}$$

The elimination rules for the dependent pair are as follows:

$$\frac{\Gamma \vdash P \in (x : A) \times B}{\Gamma \vdash \mathsf{fst}(P) \in A}\times\mathrm{E}_1 \qquad \frac{\Gamma \vdash P \in (x : A) \times B}{\Gamma \vdash \mathsf{snd}(P) \in [\mathsf{fst}(P)/x]B}\times\mathrm{E}_2$$

Proof. It suffices to validate the elimination rules in case $\Gamma \equiv \cdot$; then, by hypothesis and inversion of the meaning of membership, we have $P \Rightarrow \langle M, N\rangle$ such that $M \in A$ and $N \in [M/x]B$. By the reduction rule for $\mathsf{fst}(\langle M, N\rangle)$ and the meaning of membership, $\times\mathrm{E}_1$ is immediately evident; because reduction is confluent, we know that $[M/x]B$ is computationally equal to $[\mathsf{fst}(P)/x]B$, whence $\times\mathrm{E}_2$ becomes evident. □

4.2.1.2 Dependent Functions

The dependent function type $(x : A) \to B$ is defined analogously. First, we augment the computation system with new operators:

$$\frac{}{(x : A) \to B \Rightarrow (x : A) \to B} \qquad \frac{}{\lambda x.\, M \Rightarrow \lambda x.\, M}$$

$$\frac{F \Rightarrow \lambda x.\, M \qquad [N/x]M \Rightarrow M'}{F\,N \Rightarrow M'}$$

Next, we define the type $(x : A) \to B$ (presuppose $A\ type$ and $x : A \vdash B\ type$) by declaring that $\lambda x.\, M$ shall be a canonical member under the circumstances that $x : A \vdash M \in B$, and moreover, that $\lambda x.\, M$ and $\lambda x.\, N$ shall be equal as canonical members under the circumstances that $x : A \vdash M = N \in B$.

Just as before, the formation and introduction rules for the dependent function type are immediately evident:

$$\frac{\Gamma \vdash A\ type \qquad \Gamma, x : A \vdash B\ type}{\Gamma \vdash (x : A) \to B\ type}\to\mathrm{F} \qquad \frac{\Gamma, x : A \vdash M \in B}{\Gamma \vdash \lambda x.\, M \in (x : A) \to B}\to\mathrm{I}$$

The elimination rule is intended to be the following:

$$\frac{\Gamma \vdash F \in (x : A) \to B \qquad \Gamma \vdash M \in A}{\Gamma \vdash F\,M \in [M/x]B}\to\mathrm{E}$$

Proof. It suffices to consider the case where $\Gamma \equiv \cdot$. By hypothesis, we have that $F \Rightarrow \lambda x.\, E$ such that $x : A \vdash E \in B$; then, the reduction rule is applicable, yielding $F\,M \Rightarrow N$. By the meaning of hypothetico-general judgment, we may deduce $N \in [M/x]B$. □

4.2.2 Justifying the let Rule

Most programming languages have something called a let expression, which satisfies a rule like the following:

$$\frac{\Gamma \vdash M \in A \qquad \Gamma, x : A \vdash N \in B \qquad x \notin FV(B)}{\Gamma \vdash \mathsf{let}\ x : A = M\ \mathsf{in}\ N \in B}\ \mathit{let}$$

We may justify this rule by extending our operational semantics with a rule for the non-canonical let operator:

$$\frac{[M/x]N \Rightarrow N'}{\mathsf{let}\ x : A = M\ \mathsf{in}\ N \Rightarrow N'}$$

Then, the *let* rule is valid under the meaning explanation.

Proof. It suffices to consider the case that $\Gamma \equiv \cdot$. By the meaning of membership under hypothetico-general judgment, we have $[M/x]N \Rightarrow N'$ such that N' is a canonical member of the type $[M/x]B$. □

4.2.3 Alternative Meaning Explanations

The standard meaning explanation for type theory is called *verificationist* because the types are defined by stating how to form a canonical member (i.e. a canonical verification); in this setting, the introduction rules are evident by definition, and the elimination rules must be shown to be *locally sound* with respect to the introduction rules. This is what we have done above.

An alternative approach is to define a type by its *uses*, and have the elimination rules be evident by definition; then, the introduction rules must be shown to be *locally complete* with respect to the elimination rules. This is called the *pragmatist* meaning explanation.

Finally, following Dummett's notion of *logical harmony*, one may choose to explain the connectives by appealing to both their introduction and elimination rules, requiring that they cohere mutually through local soundness and local completeness (Pfenning 2002).

4.3 Dependent Types for Pragmatics

In Dynamic Semantics, the discourse "A man walked in. He sat down." would be represented by a proposition like the following:

$$(\exists x : \mathsf{E}.\mathit{Man}\, x \wedge \mathit{WalkedIn}\, x) \wedge \mathit{SatDown}\, x$$

In standard presentations of semantics, of course, the above would be a malformed proposition, because x is out of scope in the right conjunct, however in Dynamic Semantics, the scope of existentials is extended aritificially to make this a well-formed proposition. Following Sundholm's 1986 revelation, however, in a dependently typed setting we may assign such a sentence the following meaning:

$$(p : (x : \mathsf{E}) \times \mathit{Man}\, x \times \mathit{WalkedIn}\, x) \times \mathit{SatDown}\, (\mathsf{fst}(p))$$

Rather than modifying the behavior of existentials, which under the dependent typing discipline become pairs, we instead use a dependent pair type in place of the conjunction. Conjunctions would become pair types regardless, but by using an explicitly dependent pair, we license the right conjunct to refer to not only the propositional content of the left conjunct, but also to the *witnesses* of the existentially quantified proposition, by way of projection.

The semantics for *a*, *man*, *walked in*, and *sat down* are, in simplified form, just direct translations from the usual semantic representations:

$$\begin{aligned} [\![\mathrm{a}]\!] &\in (\mathsf{E} \to \mathsf{Set}) \to (\mathsf{E} \to \mathsf{Set}) \to \mathsf{Set} \\ [\![\mathrm{a}]\!] &= \lambda P.\, \lambda Q.\, (x : \mathsf{E}) \times P\, x \times Q\, x \end{aligned}$$

$$\begin{aligned} [\![\mathrm{man}]\!] &\in \mathsf{E} \to \mathsf{Set} \\ [\![\mathrm{man}]\!] &= \mathit{Man} \end{aligned}$$

$$\begin{aligned} [\![\mathrm{walkedin}]\!] &\in \mathsf{E} \to \mathsf{Set} \\ [\![\mathrm{walkedin}]\!] &= \mathit{WalkedIn} \end{aligned}$$

$$\begin{aligned} [\![\mathrm{satdown}]\!] &\in \mathsf{E} \to \mathsf{Set} \\ [\![\mathrm{satdown}]\!] &= \mathit{SatDown} \end{aligned}$$

Conjunction (in the form of sentence sequencing) is easily assigned a meaning in a similar way:

$$\begin{aligned} [\![\mathrm{S}_1.\mathrm{S}_2.]\!] &\in \mathsf{Set} \\ [\![\mathrm{S}_1.\ \mathrm{S}_2.]\!] &= (p : [\![\mathrm{S}_1]\!]) \times [\![\mathrm{S}_2]\!] \end{aligned}$$

But when we come to the meaning of the pronoun *he*, we run into a problem. What could it possibly be? For the example that we are currently considering, we need $[\![\text{he}]\!] = \mathsf{fst}(p)$, but this is not in general a solution for arbitrary occurrences of the pronoun, since it depends on the name and type of the free variable p.

Consider now the discourse "A man walked in. The man (then) sat down." The use of *the man* in the right conjunct, instead of *he*, introduces presuppositional content via the definite determiner. Ideally, the semantics of this should be nearly identical to those of the previous example (modulo β reduction). By giving *the* a dependently typed meaning, we can achieve this relatively simply:

$$[\![\text{the}]\!] \in (P : \mathsf{E} \to \mathsf{Set}) \to (x : \mathsf{E}) \to P\,x \to E$$
$$[\![\text{the}]\!] = \lambda P.\, \lambda x.\, \lambda q.\, x$$

The first argument to *the* is simply the predicate, which in this case will be *Man*. The second argument is an entity, and the third is an inhabitant of the type $P\,x$, i.e. a witness that $P\,x$ holds. Therefore we would want:

$$[\![\text{the man}]\!] = (\lambda P.\, \lambda x.\, \lambda q.\, x)\ (\mathit{Man}\,(\mathsf{fst}(p)))\ (\mathsf{fst}(\mathsf{snd}(p))) =_\beta \mathsf{fst}(p)$$

The term $\mathsf{fst}(p) : \mathsf{E}$ is the man referred to in the left conjunct. $\mathsf{snd}(p)$ is a witness that he is in fact a man, and that he walked in, and so $\mathsf{fst}(\mathsf{snd}(p))$ is the witness that he is a man. The argument $\mathsf{fst}(p)$ is, in effect, the solution to the presupposition induced by *the*, and $\mathsf{fst}(\mathsf{snd}(p))$ is the witness that the propositional component of the presupposition holds.

The next two pairs of examples go hand in hand. Consider the classic donkey anaphora sentences "If a farmer owns a donkey, he beats it." and "Every farmer who owns a donkey beats it." A typical Dynamic Semantics approach might assign these sentences the following meaning:

$$\forall x : \mathsf{E}.\mathit{Farmer}\,x \wedge (\exists y : \mathsf{E}.\mathit{Donkey}\,y \wedge \mathit{Owns}\,x\,y) \Rightarrow \mathit{Beats}\,x\,y$$

In the dependently typed setting, we can assign a similar meaning, but which has a more straightforward connection to the syntax (for convenience, we define the subscript p_i to project the ith element of a right nested tuple):

$$(p : (x : \mathsf{E}) \times \mathit{Farmer}\,x \times (y : \mathsf{E}) \times \mathit{Donkey}\,y \times \mathit{Owns}\,x\,y) \to \mathit{Beats}\,p_1\,p_3$$

The lexical entries for the content words and pronouns should be obvious at this point, but for *if*, *a*, and *every* we can define:

$$[\![\text{if}]\!] \in \mathsf{Set} \to \mathsf{Set} \to \mathsf{Set}$$
$$[\![\text{if}]\!] = \lambda P.\, \lambda Q.\, (p : P) \to Q$$

$$[\![\text{a}]\!] \in (\mathsf{E} \to \mathsf{Set}) \to (\mathsf{E} \to \mathsf{Set}) \to \mathsf{Set}$$
$$[\![\text{a}]\!] = \lambda P.\, \lambda Q.\, (x : \mathsf{E}) \times P\, x \times Q\, x$$

$$[\![\text{every}]\!] \in (\mathsf{E} \to \mathit{Type}) \to (\mathsf{E} \to \mathsf{Set}) \to \mathsf{Set}$$
$$[\![\text{every}]\!] = \lambda P.\, \lambda Q.\, (p : (x : \mathsf{E}) \times P\, x) \to Q(\mathsf{fst}(p))$$

With these, we can get:

$[\![\text{a farmer owns a donkey}]\!]$
$= (x : \mathsf{E}) \times \mathit{Farmer}\, x \times (y : \mathsf{E}) \times \mathit{Donkey}\, y \times \mathit{Owns}\, x\, y$

$[\![\text{if a farmer owns a donkey}]\!]$
$= \lambda Q.\, (p : (x : \mathsf{E}) \times \mathit{Farmer}\, x \times (y : \mathsf{E}) \times \mathit{Donkey}\, y \times \mathit{Owns}\, x\, y) \to Q$

$[\![\text{farmer who owns a donkey}]\!]$
$= \lambda x.\, \mathit{Farmer}\, x \times (y : \mathsf{E}) \times \mathit{Donkey}\, y \times \mathit{Owns}\, x\, y$

$[\![\text{every farmer who owns a donkey}]\!]$
$= \lambda Q.\, (p : (x : \mathsf{E}) \times \mathit{Farmer}\, x \times (y : \mathsf{E}) \times \mathit{Donkey}\, y \times \mathit{Owns}\, x\, y)$
$\quad \to Q\, p_1$

$[\![\text{beats it}]\!]$
$= \lambda z.\, \mathit{Beats}\, z\, p_3$

$[\![\text{he beats it}]\!]$
$= \mathit{Beats}\, p_1\, p_3$

We echo Sundholm's conclusion that the treatment of donkey-sentences licensed in Martin-Löf's type theory is not ad hoc, but rather is reflective of the general suitability of the framework:

> In this manner, then, the type-theoretic abstractions suffice to solve the problem of the pronominal back-reference in [the donkey-sentence]. It should be noted that there is nothing ad hoc about the treatment, since all the notions used have been introduced for mathematical reasons in complete independence of the problem posed by [the donkey-sentence]. (Sundholm 1986, p. 503)

4.3.1 Terms for Presuppositions

Provided that we can devise a general mechanism to assign the meanings given above to pronouns and definite determiners, our semantics will work just as well as standard techniques like Discourse Representation Theory or Dynamic Semantics, but in a well-scoped manner.

A number of possible solutions exist to do precisely this sort of thing in the programming languages literature. Haskell's type class constraints (Marlow 2010) and Agda's instance arguments (Devriese and Piessens 2011) provide very similar functionality but for somewhat different purposes, so one option would be to repurpose those ideas.

Haskell's type classes, however, depend on global reasoning and an anti-modular coherence condition which makes them inapplicable to our use-case, since in general there will be many solutions to a presupposition. Agda's instance arguments are closer to our needs, but we believe that a simpler approach is warranted which lends direct insight into the semantics and pragmatics of presuppositions.

The approach we will take here involves a new operator (require) that binds variables for presupposed parts of an expression. Terms, contexts and signatures are defined as follows:

$$
\begin{array}{lrcl}
\textit{Terms} & M, N, A, B & ::= & x \mid \mathsf{Set}_i \\
 & & \mid & (x : A) \to B \mid \lambda x.\, M \mid MN \\
 & & \mid & (x : A) \times B \mid \langle M, N \rangle \mid \mathsf{fst}(M) \mid \mathsf{snd}(M) \\
 & & \mid & \mathsf{require}\ x : A\ \mathsf{in}\ M \\
\textit{Contexts} & \Gamma & ::= & \cdot \mid \Gamma, x : A \\
\textit{Signatures} & \Sigma & ::= & \cdot \mid \Sigma, x : A
\end{array}
$$

The new term require $x : A$ in M should be understood to mean roughly "find some $x : A$ and make it available in M." In this version of type theory, we replace the judgment A *type* with membership in a universe, $A \in \mathsf{Set}_i$; except where ambiguous, we omit the level from a universe expression, writing Set.

Lexical constants (e.g. *Man*, *Own*, etc.) are to be contained in a *signature* Σ, whereas the context Γ is reserved for local hypotheses. The use of signatures to carry the constants of a theory originates from the Edinburgh Logical Framework, where individual logics were represented as signatures of constants which encode their syntax, judgments and rule schemes (Harper et al. 1993; Harper and Licata 2007). Then the basic forms of judgment are as follows:

$$
\begin{array}{rl}
\vdash\ \Sigma\ \mathit{sig} & \Sigma \text{ is a valid signature} \\
\vdash_\Sigma\ \Gamma\ \mathit{ctx} & \Gamma \text{ is a valid context} \\
\Gamma \vdash_\Sigma M : A & M \text{ has type } A
\end{array}
$$

In context validity judgments $\vdash_\Sigma\ \Gamma\ \mathit{ctx}$, we presuppose $\vdash\ \Sigma\ \mathit{sig}$; likewise, in typing judgments $\Gamma \vdash_\Sigma M \in A$, we presuppose $\vdash_\Sigma\ \Gamma\ \mathit{ctx}$. The rules for the signature and context validity judgments are as expected:

$$
\frac{}{\vdash \cdot\ \mathit{sig}} \qquad \frac{\vdash \Sigma\ \mathit{ctx} \qquad \cdot \vdash_\Sigma A \in \mathsf{Set} \qquad x \notin \Sigma}{\vdash \Sigma, x : A\ \mathit{sig}}
$$

$$
\frac{}{\vdash_\Sigma \cdot\ \mathit{ctx}} \qquad \frac{\vdash_\Sigma \Gamma\ \mathit{ctx} \qquad \Gamma \vdash_\Sigma A \in \mathsf{Set} \qquad x \notin \Gamma \cup \Sigma}{\vdash_\Sigma \Gamma, x : A\ \mathit{ctx}}
$$

Constants and hypotheses may be projected from signatures and contexts respectively:

$$\frac{}{\Gamma \vdash_{\Sigma,x:A,\Sigma'} x \in A}\ \text{const} \qquad \frac{}{\Gamma, x : A, \Gamma' \vdash_{\Sigma} x \in A}\ \text{hyp}$$

The inference rules for the familiar terms are the usual ones:

$$\frac{i < j}{\Gamma \vdash_{\Sigma} \mathsf{Set}_i \in \mathsf{Set}_j}\ \text{cumulativity}$$

$$\frac{\Gamma \vdash_{\Sigma} A \in \mathsf{Set} \qquad \Gamma, x : A \vdash_{\Sigma} B \in \mathsf{Set}}{\Gamma \vdash_{\Sigma} (x : A) \to B \in \mathsf{Set}}\ \to\text{F}$$

$$\frac{\Gamma, x : A \vdash_{\Sigma} M \in B}{\Gamma \vdash_{\Sigma} \lambda x.\, M \in (x : A) \to B}\ \to\text{I}$$

$$\frac{\Gamma \vdash_{\Sigma} M \in (x : A) \to B \qquad \Gamma \vdash_{\Sigma} N \in A}{\Gamma \vdash_{\Sigma} M\, N \in [N/x]B}\ \to\text{E}$$

$$\frac{\Gamma \vdash_{\Sigma} A \in \mathsf{Set} \qquad \Gamma, x : A \vdash_{\Sigma} B \in \mathsf{Set}}{\Gamma \vdash_{\Sigma} (x : A) \times B \in \mathsf{Set}}\ \times\text{F}$$

$$\frac{\Gamma \vdash_{\Sigma} M \in A \qquad \Gamma \vdash_{\Sigma} N \in [M/x]B}{\Gamma \vdash_{\Sigma} \langle M, N \rangle \in (x : A) \times B}\ \times\text{I}$$

$$\frac{\Gamma \vdash_{\Sigma} P \in (x : A) \times B}{\Gamma \vdash_{\Sigma} \mathsf{fst}(P) \in A}\ \times\text{E}_1$$

$$\frac{\Gamma \vdash_{\Sigma} P \in (x : A) \times B}{\Gamma \vdash_{\Sigma} \mathsf{snd}(P) \in [\mathsf{fst}(P)/x]B}\ \times\text{E}_2$$

The only inference rule which is new deals with presuppositions:

$$\frac{\Gamma \vdash_{\Sigma} M \in A \qquad \Gamma \vdash_{\Sigma} [M/x]N \in B \qquad x \notin FV(B)}{\Gamma \vdash_{\Sigma} \mathsf{require}\ x : A\ \mathsf{in}\ N \in B}\ \textit{require}$$

We can now provide a semantics for pronouns and definite determiners:

$$\begin{aligned} [\![\text{he}]\!] &= \mathsf{require}\ x : \mathsf{E}\ \mathsf{in}\ x \\ [\![\text{it}]\!] &= \mathsf{require}\ x : \mathsf{E}\ \mathsf{in}\ x \\ [\![\text{the}]\!] &= \lambda P.\, \mathsf{require}\ x : \mathsf{E}\ \mathsf{in}\ (\mathsf{require}\ p : P\, x\ \mathsf{in}\ x) \end{aligned}$$

Now let us reconsider our examples with the new semantics:

$$[\![\text{A man walked in. He sat down.}]\!]$$
$$= (p : (x : \mathsf{E}) \times \mathit{Man}\, x \times \mathit{WalkedIn}\, x) \times \mathit{SatDown}(\mathsf{require}\ y : \mathsf{E}\ \mathsf{in}\ y)$$
$$[\![\text{A man walked in. The man (then) sat down.}]\!]$$
$$= (p : (x : \mathsf{E}) \times \mathit{Man}\, x \times \mathit{WalkedIn}\, x)$$
$$\times \mathit{SatDown}\ (\mathsf{require}\ y : \mathsf{E}\ \mathsf{in}\ (\mathsf{require}\ q : \mathit{Man}\, y\ \mathsf{in}\ y))$$
$$[\![\text{If a farmer owns a donkey, he beats it.}]\!]$$
$$= (p : (x : \mathsf{E}) \times \mathit{Farmer}\, x \times (y : \mathsf{E}) \times \mathit{Donkey}\, y \times \mathit{Owns}\, x\, y)$$
$$\to \mathit{Beats}\ (\mathsf{require}\ z : \mathsf{E}\ \mathsf{in}\ z)\ (\mathsf{require}\ w : \mathsf{E}\ \mathsf{in}\ w)$$
$$[\![\text{Every farmer who owns a donkey beats it.}]\!]$$
$$= (p : (x : \mathsf{E}) \times \mathit{Farmer}\, x \times (y : \mathsf{E}) \times \mathit{Donkey}\, y \times \mathit{Owns}\, x\, y)$$
$$\to \mathit{Beats}\ p_1\ (\mathsf{require}\ w : \mathsf{E}\ \mathsf{in}\ w)$$

4.3.2 Computational Justification of the require Rule

A require expression is, in essence, the same as a let expression, as found in many programming languages, except that the definiens is supplied by fiat. Its formation rule is a bit strange, of course, because the presupposition's witness appears in the premises but not in the conclusion; from a type-theoretic perspective, however, this is acceptable.

For instance, many of the rules of Computational Type Theory (Allen et al. 2006; Constable et al. 1986) strategically forget their premises, yielding novel and useful constructions such as *set types* $\{x : A \mid B(x)\}$ and *squash types* $\downarrow A$. On the other hand, this causes the typing judgment to become synthetic (Martin-Löf 1994): the evidence for the judgment is not recoverable from the statement of the judgment itself, but must be constructed by the knowing subject.

The introduction of types whose members do not contain their own typing derivations is completely justified under the verificationist meaning explanation, but this does not suffice to explain the *require* rule, which is not part of the definition of a new connective. Intuitionistic validity for *require* must be established in the same way as the validity of *let*, i.e. by computation. However, it is clear that we cannot devise an effective operation which produces out of thin air a solution to an arbitrary presupposition if there is one, since this would entail deciding the truth of any proposition (and solving Turing's Halting Problem).

This, however, does not pose an obstacle for an intuitionistic justification of this rule, since assertion acts are tensed (van Atten 2007). Because evaluation itself is an

assertion, we may explain the meaning of the judgment $\mathsf{require}\ x : A\ \mathsf{in}\ N \Rightarrow N'$ by appealing to the state of knowledge at the time of assertion.

Informally, at time n, the value of $\mathsf{require}\ x : A\ \mathsf{in}\ N$ shall be, for any witness $M \in A$ that has been experienced by time n, the value of the substitution $[M/x]N$. It should be noted, then, that the computational behavior of this operator is non-deterministic, since in general the truth of A shall have been experienced in many different ways (corresponding to the number of known solutions to the presupposition).

This explanation suffices to validate the *require* rule in light of the meaning explanation which was propounded in Sect. 4.2:

$$\frac{\Gamma \vdash_\Sigma M \in A \qquad \Gamma \vdash_\Sigma [M/x]N \in B \qquad x \notin FV(B)}{\Gamma \vdash_\Sigma \mathsf{require}\ x : A\ \mathsf{in}\ N \in B}\ \textit{require}$$

Proof. It suffices to validate the rule in case $\Gamma \equiv \cdot$; then, we must show that $\mathsf{require}\ x : A\ \mathsf{in}\ N \Rightarrow N'$ such that $N' \in B$. By our definition, the require term shall have a value in case a witness for A has been experienced; but this is already the case from the hypothesis $M \in A$. By inverting the hypothesis $[M/x]N \in B$, we have $[M/x]N \Rightarrow N'$ such that N' is a canonical member of B. □

This concludes the intuitionistic justification of the *require* rule.

4.3.2.1 Discussion and Related Work

The augmentation of our computation system with a non-deterministic oracle (require) may be viewed as a computational effect. The behavior of require is defined separately at every type A, and therefore cannot be computed by a recursive algorithm; this "infinitely large" definition is acceptable in type theory because we make no a priori commitment to satisfy Church's Thesis, which states that every effective operation is recursive. Accepting the possibility of effective but non-recursive operations leads to a property called *computational open-endedness* (Howe 1991), and endows the intuitionistic continuum with the full richness of the classical one (van Atten 2007).

The explanation of the computational behavior of the require operator is related to the Brouwer's theory of the Creating Subject, and may be seen as a "proof-relevant" version of Kripke's Schema. Sundholm explains how the Kreisel-Myhill axiomatization of the Creating Subject may be treated propositionally in Martin-Löf's type theory, relative to the existence of a uniform verification object for Kripke's Schema (Sundholm 2014).

In the same way as we have exploited the intensional character of assertion acts in intuitionistic mathematics, Coquand and Jaber (2012) prove the uniform continuity principle by adding a generic element f to their computation system, representing a Cohen real; their interpretation results in a non-trivial combination of realizability with Beth/Kripke semantics.

Finally, Rahli and Bickford (2016) add two computational effects to type theory (dynamic symbol generation and exception handling), and use them to prove Brouwer's continuity theorem and justify bar induction on monotone bars.

4.3.3 Elaboration

In addition to the computational justification of *require*, we may give a proof-theoretic justification by showing how to eliminate all instances of require from a term via elaboration.[2] To this end, we will define a meta-operation $\textsc{elab}(\mathcal{D})$ which transforms a derivation $\mathcal{D} :: \Gamma \vdash M \in A$ into an elaborated term M' which is like M but with require expressions replaced by their solutions. We define the operation inductively over the structure of the derivations as follows:

$$\textsc{elab}\left(\frac{}{\Gamma \vdash_\Sigma x \in A}\,\text{const}\right) \rightsquigarrow x$$

$$\textsc{elab}\left(\frac{}{\Gamma \vdash_\Sigma x \in A}\,\text{hyp}\right) \rightsquigarrow x$$

$$\textsc{elab}\left(\frac{}{\Gamma \vdash_\Sigma \mathsf{Set}_i \in \mathsf{Set}_j}\,\text{cumulativity}\right) \rightsquigarrow \mathsf{Set}_i$$

$$\textsc{elab}\left(\frac{\overset{\mathcal{D}}{\Gamma \vdash_\Sigma A \in \mathsf{Set}} \qquad \overset{\mathcal{E}}{\Gamma, x : A \vdash_\Sigma B \in \mathsf{Set}}}{\Gamma \vdash_\Sigma (x : A) \to B \in \mathsf{Set}}\,{\to}\mathrm{F}\right) \rightsquigarrow (x : \textsc{elab}(\mathcal{D})) {\to} \textsc{elab}(\mathcal{E})$$

$$\textsc{elab}\left(\frac{\overset{\mathcal{D}}{\Gamma, x : A \vdash_\Sigma M \in B}}{\Gamma \vdash_\Sigma \lambda x.B \in (x : A) \to B}\,{\to}\mathrm{I}\right) \rightsquigarrow \lambda x.\, \textsc{elab}(\mathcal{D})$$

$$\textsc{elab}\left(\frac{\overset{\mathcal{D}}{\Gamma \vdash_\Sigma M \in (x : A) \to B} \qquad \overset{\mathcal{E}}{\Gamma \vdash_\Sigma N \in A}}{\Gamma \vdash_\Sigma M\,N \in [N/x]B}\,{\to}\mathrm{E}\right) \rightsquigarrow \textsc{elab}(\mathcal{D})\ \textsc{elab}(\mathcal{E})$$

$$\textsc{elab}\left(\frac{\overset{\mathcal{D}}{\Gamma \vdash_\Sigma A \in \mathsf{Set}} \qquad \overset{\mathcal{E}}{\Gamma, x : A \vdash_\Sigma B \in \mathsf{Set}}}{\Gamma \vdash_\Sigma (x : A) \times B \in \mathsf{Set}}\,{\times}\mathrm{F}\right) \rightsquigarrow (x : \textsc{elab}(\mathcal{D})) {\times} \textsc{elab}(\mathcal{E})$$

$$\textsc{elab}\left(\frac{\overset{\mathcal{D}}{\Gamma \vdash_\Sigma M \in A} \qquad \overset{\mathcal{E}}{\Gamma \vdash_\Sigma N \in [M/x]B}}{\Gamma \vdash_\Sigma \langle M, N\rangle \in (x : A) \times B}\,{\times}\mathrm{I}\right) \rightsquigarrow \langle \textsc{elab}(\mathcal{D})\,,\, \textsc{elab}(\mathcal{E})\rangle$$

$$\textsc{elab}\left(\frac{\overset{\mathcal{D}}{\Gamma \vdash_\Sigma P \in (x : A) \times B}}{\Gamma \vdash_\Sigma \mathsf{fst}(P) \in A}\,{\times}\mathrm{E}_1\right) \rightsquigarrow \mathsf{fst}(\textsc{elab}(\mathcal{D}))$$

[2] From a formalistic perspective, the elaboration is all that is needed to justify the rule.

$$\textsc{elab}\left(\dfrac{\begin{array}{c}\mathcal{D}\\ \Gamma \vdash_\Sigma P \in (x : A) \times B\end{array}}{\Gamma \vdash_\Sigma \mathsf{snd}(P) \in [\mathsf{fst}(P)/x]B}\ \times\mathrm{E}_2\right) \rightsquigarrow \mathsf{snd}(\textsc{elab}(\mathcal{D}))$$

$$\textsc{elab}\left(\dfrac{\begin{array}{cc}\mathcal{D} & \mathcal{E}\\ \Gamma \vdash_\Sigma M \in A & \Gamma \vdash_\Sigma [M/x]N \in B\end{array}}{\Gamma \vdash_\Sigma \mathsf{require}\ x : A\ \mathsf{in}\ N \in B}\ \text{require}\right) \rightsquigarrow \textsc{elab}(\mathcal{E})$$

The most crucial rule is the last one—the preceding ones simply define elaboration by induction on the structure of derivations other than those for require expressions. For a require expression, however, we substitute the proof of the presupposed content for the variable in the body of the require expression.

It is evident that the elaboration process preserves type.

Theorem 4.1. *Given a derivation* $\mathcal{D} :: \Gamma \vdash_\Sigma M \in A$*, there exists another derivation* $\mathcal{D}' :: \Gamma \vdash_\Sigma \textsc{elab}(\mathcal{D}) : A$.

Proof. By induction on the structure of $\mathcal{D}$. □

An example of elaboration in action is necessary, so consider again the sentence "A man walked in. He sat down." Prior to elaboration, its meaning will be:

$$(p : (x : \mathsf{E}) \times \mathit{Man}\, x \times \mathit{WalkedIn}\, x) \times \mathit{SatDown}(\mathsf{require}\ x : \mathsf{E}\ \mathsf{in}\ x)$$

Now let $\Sigma = \mathit{Man} : \mathsf{E} \to \mathsf{Set}, \mathit{WalkedIn} : \mathsf{E} \to \mathsf{Set}, \mathit{SatDown} : \mathsf{E} \to \mathsf{Set}$. After constructing a derivation that the above type is a Set under the signature Σ, we can elaborate the associated term. The left conjunct elaborates to itself, so we will not look at that, but the elaboration for the right conjunct is more interesting. The derivation for the right conjunct, letting $\Gamma = p : (x : \mathsf{E}) \times \mathit{Man}\, x \times \mathit{WalkedIn}\, x$, is:

$$\dfrac{\dfrac{}{\Gamma \vdash_\Sigma \mathit{SatDown} \in \mathsf{E} \to \mathsf{Set}}\ \text{const} \qquad \dfrac{\begin{array}{c}\mathcal{D}\\ \cdots \quad \Gamma \vdash_\Sigma \mathsf{fst}(p) \in \mathsf{E}\end{array}}{\Gamma \vdash_\Sigma \mathsf{require}\ x : \mathsf{E}\ \mathsf{in}\ x \in \mathsf{E}}\ \text{require}}{\Gamma \vdash_\Sigma \mathit{SatDown}(\mathsf{require}\ x : \mathsf{E}\ \mathsf{in}\ x) \in \mathsf{Set}}\ \to\mathrm{E}$$

Inductively, we get:

$$\textsc{elab}\left(\dfrac{}{\Gamma \vdash_\Sigma \mathit{SatDown} \in \mathsf{E} \to \mathsf{Set}}\ \text{const}\right) \rightsquigarrow \mathit{SatDown}$$

$$\textsc{elab}\left(\begin{array}{c}\mathcal{D}\\ \Gamma \vdash_\Sigma \mathsf{fst}(p) \in \mathsf{E}\end{array}\right) \rightsquigarrow \mathsf{fst}(p)$$

For the require expression's elaboration, we substitute $\mathsf{fst}(p)$ in for x in x to get the following:

$$\textsc{elab}\left(\dfrac{\begin{array}{c}\mathcal{D}\\ \cdots \quad \Gamma \vdash_\Sigma \mathsf{fst}(p) \in \mathsf{E}\end{array}}{\Gamma \vdash_\Sigma \mathsf{require}\ x : \mathsf{E}\ \mathsf{in}\ x \in \mathsf{E}}\ \text{require}\right) \rightsquigarrow \mathsf{fst}(p)$$

And finally the elaboration of whole subderivation yields $\mathit{SatDown}(\mathsf{fst}(p))$, and so the complete derivation yields

$$(p : (x : \mathsf{E}) \times \mathit{Man}\, x \times \mathit{WalkedIn}\, x) \times \mathit{SatDown}(\mathsf{fst}(p))$$

which is the meaning we had wanted.

A similar proof for "A man walked in. The man (then) sat down." can be given, with an extra non-trivial branch for $\mathit{Man}(\mathsf{fst}(p))$. Focusing just on the subproof for *the man*, we have the following typing derivation:

$$\dfrac{\dfrac{\mathcal{E}}{\Gamma \vdash_\Sigma \mathsf{fst}(p) \in \mathsf{E}} \qquad \dfrac{\dfrac{\dfrac{\dfrac{}{\Gamma \vdash_\Sigma p \in (x : \mathsf{E}) \times \mathit{Man}\, x \times \mathit{WalkedIn}\, x}\,\mathrm{hyp}}{\Gamma \vdash_\Sigma \mathsf{snd}(p) \in \mathit{Man}(\mathsf{fst}(p)) \times \mathit{WalkedIn}(\mathsf{fst}(p))}\,\times\mathrm{E}_2}{\Gamma \vdash_\Sigma \mathsf{fst}(\mathsf{snd}(p)) \in \mathit{Man}(\mathsf{fst}(p))}\,\times\mathrm{E}_1 \qquad \dfrac{\mathcal{E}}{\Gamma \vdash_\Sigma \mathsf{fst}(p) \in \mathsf{E}}}{\Gamma \vdash_\Sigma \mathsf{require}\ q : \mathit{Man}(\mathsf{fst}(p))\ \mathsf{in}\ \mathsf{fst}(p) \in \mathsf{E}}\,\mathrm{require}}{\Gamma \vdash_\Sigma \mathsf{require}\ x : \mathsf{E}\ \mathsf{in}\ (\mathsf{require}\ q : \mathit{Man}\, x\ \mathsf{in}\ x) \in \mathsf{E}}\,\mathrm{require}$$

This similarly elaborates to $\mathsf{fst}(p)$ just as the subproof for *he* did before.

Elaboration for "If a farmer owns a donkey, he beats it." and "Every farmer who owns a donkey beats it." unfolds in a similar fashion, with the elaboration of the antecedent $(x : \mathsf{E}) \times \mathit{Farmer}\, x \times (y : \mathsf{E}) \times \mathit{Donkey}\, y \times \mathit{Owns}\, x\, y$ being trivial. The consequent *Beats* (require $z : \mathsf{E}$ in z) (require $w : \mathsf{E}$ in w) breaks down into three subproofs, one for the predicate *Beats* which elaborates trivially, and the two require subproofs which elaborate like the previous pronominal examples. The only difference now is that the context licenses more options for the proofs.

Keen eyes will notice, however, that there should be four solutions, because both require expressions demand something of type E—the words *he* and *it* have no gender distinction in the semantics. This is left as an unspecified part of the framework, as there are a number of options for resolving gender constraints. Two options that are immediately obvious are (1) make E itself a primitive function $\mathsf{E} : \mathit{Gender} \to \mathsf{Set}$ and then specify a gender appropriately, or (2) add another require expression so that, for example, $[\![\text{he}]\!] = \mathsf{require}\ x : \mathsf{E}\ \mathsf{in}\ (\mathsf{require}\ p : \mathit{Masc}\, x\ \mathsf{in}\ x)$ and provide appropriate axioms (possibly simply by deferring to other cognitive systems for judging gender). The former solution is akin to how certain versions of HPSG treat gender as a property of indices not of syntactic elements.

4.4 Discussion and Related Work

In the previous sections, we have described an approach to pronominal and presuppositional pragmatics based on dependent types, as an alternative to DRT and Dynamic Semantics. The main difference from a standard dependently typed λ calculus is the addition of require expressions, and an elaboration process to eliminate them.

4.4.1 Contextual Modal Type Theory

Another approach would be to eliminate require expressions by adopting Contextual Modal Type Theory (CMTT) to support metavariables for presuppositions (Nanevski et al. 2008). From our perspective, our system relates to CMTT in the same way that programming directly with computational effects relates to programming in a monad. Indeed, the intuitionistic justification of the *require* rule obtains by adding an intensional effect to our computation system, whereas a CMTT-based solution would involve solving presuppositions after the fact by providing a substitution.

A modal extension can also provide an interesting solution to another well-known problem in pragmatics. Consider the sentence "John will pull the rabbit out of the hat" when said of a scene that has three rabbits, three hats, but only a single rabbit in a hat. This sentence seems to be pragmatically acceptable and unambiguous, despite there being neither a unique rabbit nor a unique hat. In the framework given above, there should be nine possible ways of resolving the presuppositions, leading to pragmatic ambiguity. A simple modality (approximately a possibility modality), however, can make sense of this: if the assertion of such a sentence presupposes that the sentence can be true via a modality (i.e. to assert P is to presuppose $\diamond P$), then there is only one way to solve the rabbit and hat presuppositions which would also make it possible to resolve the possibility presupposition—pick the rabbit that is in a hat, and the hat that the rabbit is in—yielding a unique, unambiguous meaning. Whether this belongs in the semantics-pragmatics or in some higher system (such as a Gricean pragmatics) is debatable, but that such a simple solution is readily forthcoming at all speaks to the power of the above framework.

4.4.2 Ranta's Type-Theoretical Grammar

The most representative use of dependent types in linguistics is Aarne Ranta's work on type-theoretical grammar (Ranta 1994), where pronominal meaning is given via inference rules for each particular pronoun or other presuppositional form. For example, the pronoun "he" can be explained by giving the following rules:

$$\frac{a : man}{he(a) : man} \qquad \frac{a : man}{he(a) = a : man}$$

The first is a typing rule, and the latter is the associated equality rule which reflects computation. This approach can generalize to any sort of presuppositional content, but leaves the question of the meaning of such expressions somewhat unanswered, since these interpretations presuppose that we have already understood the solution to the presupposition.

A discourse context without any possible antecedent will not merely cause a *type membership* error, as in the system presented in this paper, but will instead not have a meaning at all, as no term can be produced. We consider this an undesirable property in a semantic formalism. Interlocutors will typically not fail to understand sentences with unknown antecedents. For example, when presented with just the sentence "he's tall" out of the blue, most people will respond by asking "who's tall?", rather than by failing to find a meaning at all. To capture this, it's necessary for the sentence to have a meaning—that is, a term produced by the parser—even in the absence of that meaning computing to a value which the listener shall judge to be a canonical proposition.

In practice, in order to give meanings to anaphora which do not presuppose knowledge of their antecedents, such a theory must be extended with selection operators, such as Bekki's @-operator (Bekki 2014) or our require operator. This technique, of separating the assignment of meanings from the assertion that they are propositional, is based directly upon Martin-Löf's reconstruction of propositional well-formedness as a judgment, rather than a mere matter of grammar (Martin-Löf 1996).

4.4.3 Bekki's @-Operator

To assign meanings to anaphora, Bekki (2014) pursued an approach similar to ours, in which an oracle operator $(@_i : A)$ was added with the following formation rule:

$$\frac{A\ \mathit{type} \qquad A\ \mathit{true}}{(@_i : A) \in A}$$

The index i allows an expression to share a presupposition with another, which is a very useful extension that might be added to our framework. Following our computational interpretation of require, we see our operator as essentially a call-by-value analogue to Bekki's $(@_i : A)$, since in require $x : A$ in N, the presupposition $x : A$ must be resolved before N shall be reduced.

We believe that our require operator is suggestive of the interactive nature of presupposition resolution; indeed, it is possible to see require $x : A$ in N as a dialogue, in which one party requests an A to fill the hole in N—and so it seems likely that the oracle's choice of a felicitous $M \in A$ shall be based in part on the sense of the intended construction $N(x)$, and we may recover the form $(@_0 : A)$ as the special case require $x : A$ in x.

Acknowledgements The second author thanks Mark Bickford, Bob Harper and Bob Constable for illuminating discussions on choice sequences, Church's Thesis, and computational open-endedness. We thank our reviewers for their constructive feedback and references to related work.

References

Allen, S., Bickford, M., Constable, R., Eaton, R., Kreitz, C., Lorigo, L., Moran, E.: Innovations in computational type theory using Nuprl. J. Appl. Log. **4**(4), 428–469 (2006). Towards Computer Aided Mathematics

Bekki, D.: Representing anaphora with dependent types. In: Asher, N., Soloviev, S. (eds.) Logical Aspects of Computational Linguistics. Lecture Notes in Computer Science, vol. 8535, pp. 14–29. Springer, Berlin/Heidelberg (2014)

Constable, R.L., Allen, S.F., Bromley, H.M., Cleaveland, W.R., Cremer, J.F., Harper, R.W., Howe, D.J., Knoblock, T.B., Mendler, N.P., Panangaden, P., Sasaki, J.T., Smith, S.F.: Implementing Mathematics with the Nuprl Proof Development System. Prentice-Hall, Upper Saddle River (1986)

Coquand, T., Jaber, G.: A computational interpretation of forcing in type theory. In: Dybjer, P., Lindström, S., Palmgren, E., Sundholm, G. (eds.) Epistemology Versus Ontology, Logic, Epistemology, and the Unity of Science, vol. 27, pp. 203–213. Springer, Dordrecht (2012)

Devriese, D., Piessens, F.: On the bright side of type classes: instance arguments in Agda. In: Proceedings of the 16th ACM SIGPLAN International Conference on Functional Programming, ICFP'11, pp. 143–155. ACM, New York (2011)

Harper, R., Licata, D.: Mechanizing metatheory in a logical framework. J. Funct. Program. **17**(4–5), 613–673 (2007)

Harper, R., Honsell, F., Plotkin, G.: A framework for defining logics. J. ACM **40**(1), 143–184 (1993)

Howe, D.: On computational open-endedness in Martin-Löf's type theory. In: Proceedings of Sixth Annual IEEE Symposium on Logic in Computer Science (LICS'91), Amsterdam, pp. 162–172 (1991)

Marlow, S.: Haskell 2010 Language Report (2010). https://www.haskell.org/definition/haskell2010.pdf

Martin-Löf, P.: Constructive mathematics and computer programming. In: Cohen, L.J., Łoś, J., Pfeiffer, H., Podewski, K.P. (eds.) Logic, Methodology and Philosophy of Science VI. Proceedings of the Sixth International Congress of Logic, Methodology and Philosophy of Science, Hannover 1979. Studies in Logic and the Foundations of Mathematics, vol. 104, pp. 153–175. North-Holland (1982)

Martin-Löf, P.: Intuitionistic Type Theory. Bibliopolis, Napoli (1984)

Martin-Löf, P.: Analytic and synthetic judgements in type theory. In: Parrini, P. (ed.) Kant and Contemporary Epistemology. The University of Western Ontario Series in Philosophy of Science, vol. 54, pp. 87–99. Springer, Dordrecht (1994)

Martin-Löf, P.: On the meanings of the logical constants and the justifications of the logical laws. Nord. J. Philos. Log. **1**(1), 11–60 (1996)

Nanevski, A., Pfenning, F., Pientka, B.: Contextual modal type theory. ACM Trans. Comput. Log. **9**(3), 23:1–23:49 (2008)

Pfenning, F.: Logical frameworks—a brief introduction. In: Schwichtenberg, H., Steinbrüggen, R. (eds.) Proof and System-Reliability. NATO Science Series, vol. 62, pp. 137–166. Springer, Dordrecht (2002)

Rahli, V., Bickford, M.: A nominal exploration of intuitionism. In: Proceedings of the 5th ACM SIGPLAN Conference on Certified Programs and Proofs (2016)

Ranta, A.: Type-Theoretical Grammar. Oxford University Press, Oxford (1994)

Sundholm, G.: Proof theory and meaning. In: Gabbay, D., Guenthner, F. (eds.) Handbook of Philosophical Logic. Synthese Library, vol. 166, pp. 471–506. Springer, Dordrecht (1986)

Sundholm, G.: Constructive recursive functions, Church's thesis, and Brouwer's theory of the creating subject: afterthoughts on a parisian joint session. In: Dubucs, J., Bourdeau, M. (eds.) Constructivity and Computability in Historical and Philosophical Perspective, Logic, Epistemology, and the Unity of Science, vol. 34, pp. 1–35. Springer, Dordrecht (2014)

van Atten, M.: Brouwer Meets Husserl: On the Phenomenology of Choice Sequences. Springer, Dordrecht (2007)

Chapter 5
On the Computational Meaning of Axioms

Alberto Naibo, Mattia Petrolo, and Thomas Seiller

Abstract This paper investigates an anti-realist theory of meaning suitable for both logical and proper axioms. Unlike other anti-realist accounts such as Dummett–Prawitz verificationism, the standard framework of classical logic is not called into question. This account also admits semantic features beyond the inferential ones: computational aspects play an essential role in the determination of meaning. To deal with these computational aspects, a relaxation of syntax is necessary. This leads to a general kind of proof theory, where the objects of study are not typed objects like deductions, but rather untyped ones, in which formulas are replaced by geometrical configurations.

Keywords Axiomatic theories • Classical logic • Anti-realist semantics • Untyped proof theory • Proof-search • Proof reduction

5.1 Introduction

5.1.1 Between Models and Proofs: The Standard Conception of Axioms

In the standard conception of axioms, the notion of *structure* has a conceptual and ontological priority. Starting from a certain « body of facts [*Tatsachenmaterial*] » (see Hilbert (1905), translated in Hallett 1995, p. 136) composed by propositions, theorems, conjectures, and proof methods belonging to different mathematical systems, it is possible to single out some *invariants* that allow to

A. Naibo (✉) • M. Petrolo
IHPST (UMR 8590), Université Paris 1 Panthéon-Sorbonne, CNRS, ENS
13 rue du Four, 75006 Paris, France
e-mail: alberto.naibo@univ-paris1.fr; mattia.petrolo@univ-paris1.fr

T. Seiller
Department of Computer Science, University of Copenhagen, Njalsgade 128–132,
Building 24, 5th floor, DK-2300 Copenhagen S, Denmark
e-mail: seiller@di.ku.dk

J. Redmond et al. (eds.), *Epistemology, Knowledge and the Impact of Interaction*,
Logic, Epistemology, and the Unity of Science 38, DOI 10.1007/978-3-319-26506-3_5

identify the common features of these systems. In this process of *abstraction*, a general and univocal form is pointed out. This form, that « might be called a relational structure » (Bernays 1967, p. 497) is fixed at the linguistic level by the axioms. When formalized in a set-theoretical way, this notion of structure becomes ontologically concrete and can play the role of an *interpretation structure* – i.e. a model – both for the axioms and for the sentences derivable from them. In this sense, we can say that the grounding idea of the axiomatic method is to capture a class of models sharing some relevant properties that distinguish them from other classes of models. An immediate consequence is that the proper axioms of a certain theory $\mathscr{T}$ are considered meaningful because they are *true* exactly in those classes of models that they identify. It seems then that the notion of axiom fits well with a truth-conditional, or model-based, theory of meaning (see Naibo 2013, ch. 3).[1] For instance, Hintikka's remark that the genuine relation between axioms and theorems is the model-theoretical relation of logical consequence, rather than the syntactical relation of derivability goes in this direction (Hintikka 2011, pp. 73–75). Derivations are then subordinated to semantical aspects, in the sense that their role is reduced to that of guaranteeing truth transmission (see Dummett 1973a, p. 434). This idea finds a further confirmation in the difficulty of constructing an inferentialist theory of meaning – for example, in the style of Dummett-Prawitz verificationism – when it has to deal with axioms. In particular, in the presence of proper axioms, the fundamental notion of *canonical proof* is lost. Consider, for example, the derivation in natural deduction of the sentence $\forall x(x = 0 \vee \exists y(x = s(y)))$ from Peano's axioms (regardless of whether we are using classical or intuitionistic inference rules). The derivation terminates with an application of the $\rightarrow$ elimination rule having as a major premiss an instance of the axiom scheme of induction. This is not a *canonical proof* in the sense of the (inferential) verificationism of Dummett-Prawitz, because it does not terminate with the introduction rule of the principal connective of the sentence under analysis – in this case the $\forall$ introduction rule.[2] In general, this means that in the presence of proper axioms it is not possible to reduce to a common form – or to identify a common feature of – all possible proofs of the sentences that have the same principal connective. The immediate consequence is that the notion of proof cannot be used to explain the meaning of sentences: in absence of a common form to which they can be reduced, different proofs of the same sentence would turn out to

[1] As Dummett (1976) remarks, a truth-conditional theory of meaning is itself presented in an axiomatic way. In particular, the axioms fix the reference of the primitive terms of the language.

[2] Notice that here we prefer to exclude from the set of canonical proofs those that are *trivially canonical*, i.e. those terminating with a sequence of $\circ$-elim/$\circ$-intro rules, with $\circ$ as the principal connective of the conclusion. Another well known example of axiomatic theories that prevent from the possibility of obtaining canonical proofs, namely of canonical proofs of disjunctive or existential sentences, is represented by those theories the axioms of which contain strictly positive occurrences of disjunctive and existential sentences (see Troelstra and Schwichtenberg 2000, pp. 6, 106–107).

confer different meanings to it, so we could not refer to it as the *same sentence*.[3] This would be particularly problematic for mathematics, where a theorem is supposed to always possess the same meaning, even if proved in different ways.

However, an inferentialist theory of meaning resting on the notion of proof seems to be particularly attractive in the case of mathematical theories, since in mathematical practice proofs are usually reckoned as a privileged way to access to mathematical objects (especially when proofs are considered as constructions) and to the properties of these objects (especially when proofs are considered as demonstrations).[4] This position has been in fact endorsed also by some champions of the axiomatic method, like the Bourbaki group, who opened its seminal book on set theory with these words:

> Ever since the time of the Greeks, mathematics has involved proof; and it is even doubted by some whether proof, in the precise and rigorous sense which the Greeks gave to this word, is to be found outside mathematics. (Bourbaki 1968, p. 7)

Axioms represent then the meeting point of different moments of the development of mathematical theories, or of the mathematical enterprise in general. On the one hand, the process of abstraction leading to the definition of an axiomatic system is connected to a *synthetic* moment: the axioms are required to capture all the relevant information belonging to a certain domain of discourse, in the sense that they should compactify and synthesize everything we know about a certain domain.[5] On the other hand, the *analytical* moment of the mathematical enterprise is represented by proofs: the information present in the axioms should be extracted and deployed just by the use of pure logical derivations (see Pasch 1925, pp. 194–195; Hempel 1945, p. 7).

Axioms have thus a double role: they are the points of entrance of the semantics into the syntax (i.e. they single out a class of models) and the starting points of derivations. This distinctive feature of axioms of connecting semantics and syntax takes its formal characterization in the *soundness and completeness theorem* (for Hilbert-style deductive systems):

$$ax_{\mathscr{T}} \vDash A \text{ if and only if } ax_{\mathscr{T}} \vdash A$$

where $ax_{\mathscr{T}}$ is the set of axioms of a certain theory $\mathscr{T}$, and A is a sentence of the language of $\mathscr{T}$.

[3] This position is usually identified with a Wittgensteinian position (see Wittgenstein 1956, Part II, §31, Part V, §7), nevertheless we think that this is a mistake. The point is that Wittgenstein is not speaking of the meaning of a sentence considering it as something abstract, invariant and objective – as the propositional content of that sentence could be – but he is speaking instead of how each single agent gets to a specific understanding of that sentence, depending on the particular place she assigns to it inside her own web of beliefs.

[4] The characterization of proofs either as constructions – or better, as methods of construction – or as demonstrations can be originally found in Proclus (1970, p. 157ff.). A more contemporary discussion about these distinctions can be found in Sundholm (1993, 1998).

[5] This aspect is strictly connected to the ideal of deductive completeness (cf. Awodey and Reck 2002, p. 4).

In this sense, even if from the axiomatic point of view the most fundamental relation between axioms and theorems are still that of logical consequence, which is in fact a relation that allows to express results holding between the sentences of the theory. However, if we do not simply want to present already established results (if we do not simply want to present these relationships between the sentences of the theory), but we also want also to explain how they have been obtained, or even to discover new ones, it seems that the notion of proof naturally plays a fundamental role. In particular, when proofs are considered, it becomes clear why in logic a central place is reserved to the soundness and completeness theorem: it allows to link the truth of sentences – supposed to be guaranteed by some kind of (abstract) structures – with the way in which it can be achieved by human agents, that is via proofs. In other words, when the soundness and completeness theorem is analyzed with respect to proofs, and not only with respect to provability,[6] it seems to open the way to an epistemic interpretation of semantical concepts otherwise transcending a human-based dimension. But then, why not to let proofs play a genuine and autonomous semantic role?[7] Do we really need (abstract) structures in order to define semantical concepts – like those of truth and meaning (of mathematical sentences) – or is possible to recover them in some more ontological parsimonious way?

5.1.2 Our Proposal: A Proof-Based Account of Axiomatic Theories

What we try to investigate here is a way to take into account a standard presentation of mathematical theories based on axioms in which proofs (and operations on them) are the only genuine semantical objects, so that there is no need to postulate other notions – like that of (set-theoretic) structure – which are highly problematic to define, since they seem to invoke a reference to some kind of abstract objects, which differently from proofs are not immediately accessible to human agents.[8] In this sense, inferentialist theories of meaning are closely related to anti-realist positions according to which semantical concepts must not to transcend our epistemic capacities. Now, if in logic and mathematics we should not abandon

[6]It is worth noting that the fact of not limiting to the simple analysis of the provability level, but to investigate theorems also from the point of view of the structural analysis of proofs is one of the leitmotifs of Kreisel's work (see for example Kreisel 1987, p. 399).

[7]This seems to be indeed the idea expressed by Bourbaki (1994, p. 17) when they say: « "Mathematical truth" resides thus uniquely in logical deduction starting from premises arbitrarily set by axioms.»

[8]This immediacy corresponds to the idea that proofs are *epistemic transparent* objects (Usberti 1997, p. 535): it is not possible that something is a proof without the possibility (for a human agent) to recognize it as such (cf. Kreisel 1962, p. 202).

this epistemic-based perspective, then we have not to abandon the semantical key concept of canonical proof, since a canonical proof is a finite object the nature of which does not go beyond our epistemic capacities. In order not to abandon this key concept, it seems that the only possible solution is to give up the notion of axiom in favor of other alternative notions, namely that of *non-logical rule of inference* (Negri and von Plato 1998) or that of *rewrite rule* (Dowek et al. 2003). However, from a philosophical point of view, this way of proceeding eventually leads to a substantially revisionist position. Indeed, embracing an inferentialist and anti-realist theory of meaning not only leads to revisionism about logical constants (by abandoning classical operators for intuitionistic ones; see e.g. Dummett (1991), pp. 291–300), but it also leads to revisionism about the commonly accepted view of mathematical theories, namely by abandoning the standard conception of a theory as a set of axioms and replacing it with the conception of a theory as a set of postulates (i.e. hypothetical actions or *Erzeugungsprinzipien*; see e.g. von Plato 2007, p. 199) or as an algorithm (i.e. a set of computational instructions, see Dowek 2010). These solutions are analyzed in details in Naibo (2013, Part III).

What we propose in this paper is a way to save a theory of meaning essentially based on the notion of inference and proof, but without necessarily abandoning the notion of axiom. In order to do that, we will adopt what can be called an *interactional* point of view. In contrast to standard Dummettian inferentialism, we will extend our set of key semantic concepts by accepting not only objects that exclusively contain correct instantiations of axioms and rules – as in the case of canonical proofs – but also objects that contain incorrect ones. For this reason, such objects will be called *paraproofs*. Differently from proofs, they are essentially *untyped* objects. This means that types – i.e. propositions or sentences – are no longer conceived as the primitive entities on which the inference rules act, but become the outcome of the interaction between paraproofs. A quite natural setting to model this notion of interaction is the *computational* one and especially the so-called Curry-Howard correspondence will be our starting point (see Sørensen and Urzyczyn 2006 for a comprehensive presentation). From such of a perspective, a formal correlation between proofs and programs is established; in particular, proofs can be seen as the surface linguistic "description" of the inner intensional behavior of programs. Our idea is then to show how it is possible to construct semantical aspects starting from the behavior of programs. More precisely, our aim is to show that knowing the meaning of an axiom does not consist in knowing which objects and structures make it true, but in knowing what is the computational behavior of the program associated to it, once the axiom in question has been put in interaction with other programs.

In a nutshell, we put forward an inferentialist approach, alternative to the usual Dummett-Prawitz verificationism. Our account is compatible with the viewpoint of classical logicians, in particular those who wish to remain (ontologically) parsimonious in the definition of semantical notions like meaning and truth.

5.2 Axioms and Computation

From a historical point of view, the proofs-as-programs correspondence was first established between (deductive systems for) constructive logics and abstract functional programming languages, and particularly between minimal or intuitionistic natural deduction and λ-calculus. In this setting, the execution of a program – i.e. a λ-term – having specification A corresponds to the normalization of a natural deduction proof having A as conclusion. More precisely, each (local) elimination of a *detour* taking the form of a $\circ$-introductio/$\circ$-elimation sequence – where $\circ$ is a logical connective – corresponds to a step of program execution. For example, the elimination of a $\rightarrow$ *detour* corresponds to the execution of one step of β-reduction, that is, one step of the computation of the value taken by the function/program $\lambda x.t$ once the latter is applied to the input u:

$$\dfrac{\dfrac{\begin{array}{c}[x:A]^{(m.)}\\ \vdots\\ t:B\end{array}}{\lambda x.t:A\rightarrow B}\rightarrow\text{intro} \qquad \begin{array}{c}\vdots\\ u:A\end{array}}{(\lambda x.t)u:B}\rightarrow\text{elim}\,(m.) \quad \rightsquigarrow \quad \begin{array}{c}\vdots\\ u:A\\ \vdots\\ t[u/x]:B\end{array}$$

Computation seems then to be necessarily tied to non-atomic – i.e. complex – types, namely the maximal formulas of the *detours*.[9] In this respect, it is worth noting that there are two types of formulas that can never appear as maximal formulas: proper axioms and assumptions.[10] The reason is trivial. Proper axioms and assumptions are always the starting points of derivations, therefore they cannot be preceded by an introduction rule and thus no detour can be created.

The analysis just sketched can be made more precise by appealing to two properties resulting from a generalization of Prawitz's translation of natural deduction into sequent calculus (Prawitz 1965, pp. 90–91; von Plato 2003, § 5) . Such generalization allow to work not only with normal proofs, but also with non-normal ones (for details see Pereira 1982, Part C).[11] We will work with this

[9]For the notion of maximal formula see Dummett (1977, p. 152). For the notation of λ-calculus see Krivine (1993).

[10]Notice that the difference between proper axioms and assumptions is that the former can never be discharged, while the latter are in principle always dischargeable (even if de facto they are not). At the level of proof-objects – i.e. at the level of the objects used for codifying derivational steps, as λ-terms (see Sundholm 1998, pp. 196–197) – the difference is that proper axioms correspond to proof-term constants, while assumptions to proof-term variables. Roughly speaking, proper axioms are sentences which are to be considered as already having been proved, and therefore which can always be justified (see Heyting 1962, p. 239). Assumptions, on the other hand, are placeholders: they wait to be justified by a proof that we neither possess nor know to be constructible.

[11]The original Prawitz' translation works for systems of minimal, intuitionistic and classical logic. It is worth noting that Prawitz treats sequents as composed by sets of formulas. However, his translation can be adapted to the case of sequents considered as composed by multisets. In this

translation because it is faithful from the computational point of view: *only* detours are translated into instances of the cut rule. In this way, cut-formulas coincide with maximal formulas and thus cuts are always non-atomic.[12] Furthermore, Prawitz' translation operates by transforming proper axioms – or their instances, in the case of axiom schemes – into part of the derivation, i.e. by moving proper axioms – or their instances – from the top position in a natural deduction derivation to the lefthand side of sequents. For example, consider the theory of equality presented by the two axioms

$$\textbf{(Ref)} \quad \forall x(x = x)$$

$$\textbf{(Euc)} \quad \forall x \forall y \forall z(x = y \wedge x = z \rightarrow y = z)$$

The non-normal proof in natural deduction shown in Fig. 5.1a is translated as the sequent calculus proof shown in Fig. 5.1b.

The immediate consequence of the translation is that two types of formulas are excluded from the set of cut-formulas:

1. The formulas that are proper axioms.
2. The principal formulas of logical axioms (Negri and von Plato 2001, p. 16), also called *identity axioms* (Girard et al. 1989, p. 30).

And since a fundamental property of any "good" sequent calculus system is the possibility of "atomizing" the principal formulas of identity axioms (Wansing 2000, pp. 10–11),[13] we can replace (2) by

2*. All the atomic formulas appearing in the logical (i.e. identity) axioms.

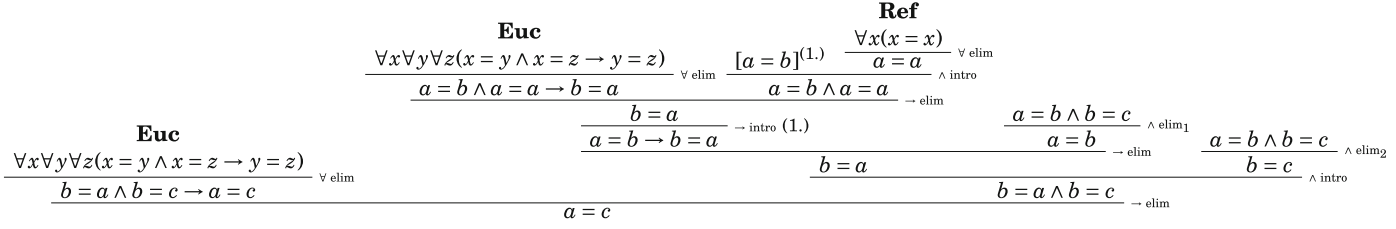
This means that from the computational point of view, proper axioms and (atomic) identity axioms are identified: neither of them plays any role in the execution of a program.[14] They have no genuine computational content, as they are just the external

case, the translation is directed either to the sequent calculus system **G1[mic]** or **G2[mic]** (see Troelstra and Schwichtenberg (2000), pp. 61–66 for a presentation of these systems).

[12]In Gentzen's translation (Genzten 1934–1935, § 4), differently from the translation chosen here, normal proofs are also translated into proofs with non-atomic cuts, because elimination rules are translated by appealing to cuts.

[13]By the translation provided above we can assign a well defined computational content to cut elimination, i.e., β-reduction. Analogously, the property of identity axiom atomization can be assigned a computational operation, i.e., η-expansion. This operation guarantees the possibility of working in an extensional setting even in the case of programs, which are by definition intensional objects (see Hindley and Seldin 2008, pp. 76–77). For further details see Naibo and Petrolo (2015).

[14]The λ-term associated to the previous natural deduction proof is $(\mathtt{t})\langle(\lambda x(\mathtt{t})\langle x, \mathtt{r}\rangle)\pi_1(z), \pi_2(z)\rangle$, where $\mathtt{r}$ and $\mathtt{t}$ are two proof-constants associated with the reflexivity and Euclidean axioms respectively, and $\pi_1(z)$ and $\pi_2(z)$ are the first and second projection of z. Reducing the β-*redex* contained in this λ-term – which corresponds to the detour of the proof that this λ-term codifies – we get $(\mathtt{t})\langle(\mathtt{t})\langle\pi_1(z), \mathtt{r}\rangle, \pi_2(z)\rangle$. It is not difficult to see that the constants $\mathtt{t}$ and $\mathtt{r}$, as well as the variable z, are not involved in the process of reduction. This means that proof-constants have no interaction with the other proof-constructors and that we cannot assign to them any computational



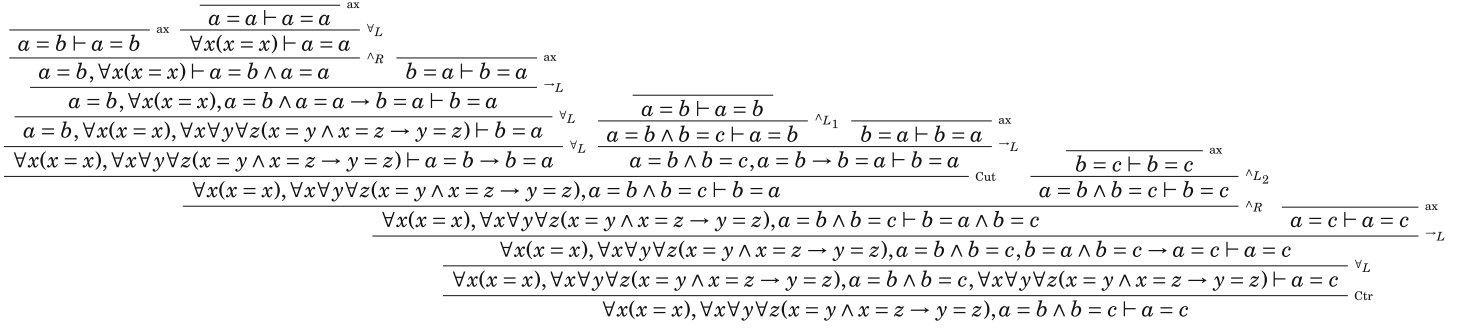


Fig. 5.1 Example of translation from natural deduction to sequent calculus. (**a**) Natural deduction proof. (**b**) Sequent calculus proof

borders of proofs. In other words, the *dynamics* of proofs, when conceived of as cut elimination, do not tell us much about the role played by both proper and identity axioms in formal proofs. But what if we look at other dynamical aspects of proofs such as proof-search procedures (i.e. bottom-up proof reconstructions)? Do these procedures give us more information about the proof-theoretical role of axioms?

First, it should be noticed that by shifting the attention from cut-elimination procedures – seen as a program executions – to proof-search procedures also entails a shift of logical framework, from intuitionistic logic – or more generally constructive logics – to classical logic. The reason is that classical systems are much more suitable to proof-search than intuitionistic ones, because of the invertibility of all their logical rules (Troelstra and Schwichtenberg 2000, p. 79). Let us then concentrate on classical sequent calculus and consider a provable sequent of the form $\Gamma, A \vdash \Delta$, where A is a proper axiom or an instance of a proper axiom scheme. If we have an algorithmic procedure allowing to reconstruct the proof of the sequent, for example by working with a system of classical logic like **G3c**, the best result we can get is to decompose A into atomic sentences belonging to some initial identity axioms.[15] Again, we should conclude that proper axioms have no particular role in proofs: they are not different from other context formulas used in purely logical proofs, and everything can eventually be reduced to logical combinations of identity axioms. In order to prevent the transformation of proofs containing proper axioms into purely logical proofs, we have to block the logical decomposition of the axiom A. A tentative strategy would be to apply a proof-search procedure on $\Gamma, A \vdash \Delta$ within a system without left-rules. This strategy is equivalent to a proof-search procedure in a right-handed system with an additional initial sequent $\vdash A$. However, in such a system, every proof using $\vdash A$ uses cuts that cannot be eliminated (Girard 1987a, p. 125; Troelstra and Schwichtenberg 2000, p. 127). In general, cuts are an obstacle to the root-first reconstruction of proof; in order to determine whether a sequent $\Gamma \vdash \Delta$ is derivable by using a cut rule, we should check the derivability of the two sequents $\Gamma \vdash C$ and $C, \Gamma \vdash \Delta$ for any arbitrary formula C, which produces the immediate consequence of removing any bound on the proof-search. Therefore, a proof-search procedure that allows the recognition of all the theorems of the theory $\mathscr{T}$ containing the axiom A does not always terminate.

As a solution to this problem, we could still operate a proof-search on a given theorem B belonging to a certain theory $\mathscr{T}$ without requiring that the derivation closes – that is, without necessarily using the axioms of $\mathscr{T}$ as the initial sequent of the form $\vdash A$. More precisely, suppose that B is a formula such that there are no positive occurrences of existential formulas and no negative occurrences

content. Proof-objects in this case have only the role of codifying the structure of the proofs to which they are associated with, but they cannot be interpreted as programs.

[15]This point becomes even clearer when applied to proof-objects . While in natural deduction the proof-objects associated with proper axioms are constants, in sequent calculus they are complex λ-terms not containing any constant. This is because in sequent calculus proper axioms are constructed in the context of derivations, i.e. in the antecedent.

of universal formulas.[16] When we work in **G3c**, a sequent $\vdash B$ having an empty antecedent can always be univocally decomposed in a set of *basic sequents* of the form

$$\bot, \ldots, \bot, P_1, \ldots, P_m \vdash Q_1, \ldots, Q_n, \bot, \ldots, \bot \tag{5.1}$$

where P_i and Q_j are atoms (Negri and von Plato 2001, p. 50).

If B is a non-logical theorem, or even an axiom, then its decomposition leads to a (possibly infinite) set of « basic mathematical sequents » (Gentzen 1938, p. 257) containing at least one sequent of the form

$$P_1, \ldots, P_m \vdash Q_1, \ldots, Q_n, \bot, \ldots, \bot \tag{5.2}$$

where $P_i \not\equiv Q_j$ for every i and j (Negri and von Plato 2001, p. 51).

It is worth noticing that atomic identity axioms are just a particular case of (1), namely when there are no $\bot$ and $P_i \equiv Q_j$ for some i and j. This remark suggests the possibility of identifying a unique way to deal with both proper axioms and identity axioms. In the next section we propose a solution along these lines[17] by introducing a *generalized* axiom rule inspired to Girard (2001).

5.3 From Proofs to Models

In this section the attention will be focused on classical logic. This choice is not simply due to practical – if not even opportunistic – reasons related to the efficacy of classical systems over intuitionistic ones with respect to proof-search problems. There is in fact a deeper, conceptual reason that has its roots in the discussion carried out in Sect. 5.1. As we mentioned there, our aim is to do justice to a non-revisionist stance with respect to the architecture of mathematical theories, where one of the characteristic features of the standard view is precisely that the underlying logic of mathematical theories is classical logic, and not intuitionistic logic.

[16] For the standard definition of positive and negative occurrences of a formula see Troelstra and Schwichtenberg (2000, p. 6).

[17] In the next section we will restrict to one-sided sequent systems. This choice is only dictated by a wish to ease the proof analysis. Notice that the result we presented above could be adapted to one-sided systems (i.e. without any formulas on the left of sequents) by replacing the two-sided notion of *basic sequent* with the corresponding notion of one-sided basic sequent, that is $\vdash P_1, \ldots, P_n, \bot, \ldots, \bot$ where the P_i are now either atoms or negations of atoms.

5.3.1 Schütte's Completeness Proof Revisited

We will now introduce a system that allows us to study both the syntactical and the semantical role played by identity and proper axioms from the unifying perspective of Schütte's completeness proof (see Schütte 1956). This system should be considered as a general framework for carrying out an abstract and formal study of the axioms, rather than a genuine deductive system. The reason for this, as it will be explained in more detail below, is that in order to have sufficient expressive power for speaking of every possible axioms we are obliged to flirt with inconsistency (cf. Propositions 5.4 and 5.16). Despite this feature of the system, its proper logical part can be singled out through the definition of some kind of *correctness criterion*. The idea is that this system represents a compact way to simultaneously deal with a family of deductive systems: by imposing some specific restrictions on the use of the generalized axiom rule it is possible to single out one specific deductive system at a time, and to characterize it as a logical or non-logical system. In other words, what we propose here is a system for studying proofs from an *abstract* point of view, where the abstraction concerns the form taken by the initial sequents.[18]

We start by presenting the propositional case, namely the system $\mathrm{pLK}^{✠}_{R}$. We will then move to the more interesting case of first order logic. Looking at the propositional part of the system will be sufficient to grasp its main features and understand how it works. Readers who are not interested in a finer analysis of the system are recommended to skip Sect. 5.3.2.

Let $\mathscr{A}$ be a set of atomic formulas.

Definition 5.1 (Formulas and Sequents). The set of formulas is inductively defined by the following grammar

$$F := P, Q \mid \neg P \mid F \vee F \mid F \wedge F \qquad (P, Q \in \mathscr{A})$$

A sequent $\Gamma \vdash \Delta$ is an ordered pair of multisets Γ, Δ of formulas.

Definition 5.2. The system $\mathrm{pLK}^{✠}_{R}$ is defined by the rules of Fig. 5.2.

$$\frac{}{\vdash \Gamma}\ ✠^{At} \qquad \frac{\vdash \Gamma, A, B}{\vdash \Gamma, A \vee B}\ \vee \qquad \frac{\vdash \Gamma, A \qquad \vdash \Gamma, B}{\vdash \Gamma, A \wedge B}\ \wedge$$

Fig. 5.2 $\mathrm{pLK}^{✠}_{R}$ rules

[18] It must be remarked that at present there is no such discipline as *abstract proof theory* as a branch of mathematical logic, in the same sense as which there is, instead, an *abstract model theory*. Our proposal can be considered as a contribution to the attempt of defining such a discipline, which is different in nature from other attempts such as those undertaken in the categorial analysis of proofs (cf. Hyland 2002, §1).

The only proviso on the application of $✠^{At}$, the generalized axiom rule, is that the formulas appearing in the sequent $\vdash \Gamma$ (if any at all) have to be atomic formulas or negations thereof.

Proposition 5.3 (Invertibility). *The rules $\vee$ and $\wedge$ are invertible.*

Proof. See Appendix A. □

Proposition 5.4. *Every sequent can be derived in* $\mathrm{pLK}_R^{✠}$.

Proof. By induction on the number of connectives in the sequent. □

Definition 5.5. Given a proof π in $\mathrm{pLK}_R^{✠}$, $\mathfrak{L}(\pi)$ is the multiset of sequents introduced by the ✠ rules in π. $\mathfrak{L}(\pi)$ is called the *set of leaves* of π.

Lemma 5.6. *Let π and π' be derivations of $\vdash \Gamma$ in* $\mathrm{pLK}_R^{✠}$, *then $\mathfrak{L}(\pi) = \mathfrak{L}(\pi')$.*

Proof. See Appendix A. □

Remark 5.7. By Proposition 5.4 and the previous lemma we can now use the notation $\mathfrak{L}(\vdash \Gamma)$, for any sequent $\vdash \Gamma$.

Definition 5.8. A sequent $\vdash \Gamma$ is *correct* when it is atomic and there exists an atom P, such that P and $\neg P$ are both in $\vdash \Gamma$. By extension, a $✠^{At}$ rule is correct when the sequent it introduces is correct.

An *incorrect* sequent is an atomic sequent that is not correct.

Definition 5.9. The system pLK_R is obtained by replacing the $✠^{At}$ rule with a rule that introduces only correct sequents. This rule will be called logical axiom rule and noted in the following manner

$$\frac{}{\vdash \Gamma, P, \neg P}\ \mathrm{ax}$$

Proposition 5.10. *A sequent $\vdash \Gamma$ is derivable in pLK_R if and only if $\mathfrak{L}(\vdash \Gamma)$ contains only correct sequents.*

Proof. See Appendix A. □

Definition 5.11. Let $\delta : \mathscr{A} \to \{0, 1\}$ be a valuation, and $\bar{\delta}$ its extension to formulas of pLK_R.

- $\delta \vDash \Gamma$ if and only if there exists at least one $A \in \Gamma$ such that $\bar{\delta}(A) = 1$
- $\vDash \Gamma$ if and only if for all valuation δ, $\delta \vDash \Gamma$.

Lemma 5.12. *For any valuation δ, $\delta \vDash \Gamma$ if and only if for all $\vdash \Delta$ in $\mathfrak{L}(\vdash \Gamma)$, $\delta \vDash \Delta$.*

Proof. See Appendix A. □

Proposition 5.13. *$\vDash \Gamma$ if and only if all sequents in $\mathfrak{L}(\vdash \Gamma)$ are correct.*

Proof. See Appendix A. □

5.3.2 *Embedding Semantics in the Syntax*

We adapt the previous proof to the first-order case. We must take particular care of the ✠ rule and of the definition of correctness. Notice that this case is more liberal than the propositional case: no conditions are imposed on the application of the ✠ rule, which can be used at every point of the derivation.

Definition 5.14. Let $\mathrm{LK}_R^{✠}$ be the system defined by the rules of Fig. 5.3.

Definition 5.15. A derivation of $\mathrm{LK}_R^{✠}$ is a finite tree obtained from the rules of Fig. 5.3 where all leaves are closed by using a ✠ rule.

Proposition 5.16. *Every sequent* $\vdash \Gamma$ *is derivable in* $\mathrm{LK}_R^{✠}$.

Proof. Take $\vdash \Gamma$ and close the derivation by an instance of ✠ rule. □

Definition 5.17. The ✠ rule that introduces the sequent $\vdash \Gamma$ is correct if there exists a formula A such that both A and $\neg A$ are in Γ.

A derivation is correct if all its ✠ rules are correct.

Definition 5.18. A ✠ rule introducing a sequent $\vdash \Gamma$ is *admissible* if either it is correct or there exists a correct derivation of $\vdash \Gamma$ in $\mathrm{LK}_R^{✠}$.

Lemma 5.19. *If there exists a derivation* π *of* $\vdash \Gamma$ *in* $\mathrm{LK}_R^{✠}$ *such that* $\mathfrak{L}(\pi)$ *contains only admissible* ✠ *rules, then* π *can be extended to a derivation* π' *of* $\vdash \Gamma$ *such that* $\mathfrak{L}(\pi')$ *contains only correct* ✠ *rules.*

Proof. See Appendix B. □

Definition 5.20. The system LK_R is obtained by replacing the ✠ rule with a rule that introduces only correct sequents. This rule will be called logical axiom rule and noted in the following manner

$$\frac{}{\vdash \Gamma, A, \neg A}\ \mathrm{ax}$$

Theorem 5.21. *The sequent* $\vdash \Gamma$ *is derivable in* LK_R *if and only if there exists a derivation* π *of* $\vdash \Gamma$ *in* $\mathrm{LK}_R^{✠}$ *and* $\mathfrak{L}(\pi)$ *contains only admissible* ✠ *rule.*

Proof. See Appendix B. □

$$\frac{}{\vdash \Gamma}\ ✠$$

$$\frac{\vdash \Gamma, A, B}{\vdash \Gamma, A \vee B}\ \vee \qquad \frac{\vdash \Gamma, A \quad \vdash \Gamma, B}{\vdash \Gamma, A \wedge B}\ \wedge$$

$$\frac{\vdash \Gamma, A[y/x]}{\vdash \Gamma, \forall x A(x)}\ \forall \ (y \text{ fresh}) \qquad \frac{\vdash \Gamma, A(t), \exists x A(x)}{\vdash \Gamma, \exists x A(x)}\ \exists$$

Fig. 5.3 The rules of $\mathrm{LK}_R^{✠}$ sequent calculus

Definition 5.22. A ✠ rule is *simple* if the sequent $\vdash \Gamma$ it introduces contains only atoms, negations of atoms, and existential formulas.

A *simple derivation* is a derivation in which all ✠ rules are simple.

Lemma 5.23. *Let π be a derivation in $\mathrm{LK}_R^{✠}$. There exists a simple extension of π.*

Proof. See appendix B. □

Definition 5.24. Let $B = \exists x A$ be an existential formula. We define the *instances* of B to be the formulas $A[t/x]$ where t is a term.

More generally, if A is a formula and C a subformula of A, the set of *instances* of C is the set of formulas $C[t_1/x_1, \ldots, t_n/x_n]$ where $t_1, \ldots, t_n$ are terms and $x_1, \ldots, x_n$ are the bound variables of C.

Lemma 5.25. *Let π be the following derivation:*

$$\frac{}{\vdash \Gamma}\ ✠$$

If π is non-admissible, then we can find a sequence of extensions of π containing all the instances of the subformulas of Γ.

Proof. See Appendix B. □

Theorem 5.26. *Let π be a derivation in $\mathrm{LK}_R^{✠}$ of a sequent $\vdash \Gamma$ containing a non-admissible ✠ rule. Then there exists a model $\mathscr{M}$ such that $\mathscr{M} \nvDash \Gamma$.*

Proof. See Appendix B. □

Theorem 5.27. *If π is a derivation of $\vdash \Gamma$ in $\mathrm{LK}_R^{✠}$ containing only admissible ✠ rules, then for all model $\mathscr{M}$, $\mathscr{M} \vDash \Gamma$.*

Corollary 5.28. *Let π and π' be derivations in $\mathrm{LK}_R^{✠}$ of the same sequent $\vdash \Gamma$. Then π contains only admissible ✠ rules if and only if π' contains only admissible ✠ rules.*

5.3.3 Axioms and Models

Differently from what happens with the standard Schütte's completeness proof for classical logic, in the revisited proof we proposed – where $\vdash \Gamma$ is derived with non-admissible instances of ✠ rule – we cannot always conclude that $\bigvee \Gamma$ is an antilogy (i.e. a sentence which is false in every possible model).[19] In fact, $\bigvee \Gamma$ could be valid in some particular theories, namely theories whose models make true every non-admissible instance of ✠ rule used in the derivation of $\bigvee \Gamma$. From the derivation of $\bigvee \Gamma$ we can thus know more about the axioms and theorems of $\mathscr{T}$. In particular, the

[19] In other words, an antilogy is the negation of a tautology.

incorrect instances of ✠ rule correspond to a set of sequents $\mathscr{S}_1, \ldots, \mathscr{S}_n$ such that: either

(i) each $\mathscr{S}_i$ is provable in every axiomatic system containing Γ, or
(ii) $\vdash \Gamma$ is provable in every axiomatic system containing $\mathscr{S}_1, \ldots, \mathscr{S}_n$.

This means that the proof-search in the system $\mathrm{LK}_R^{✠}$ does not always constitute a method for invalidating non-logical sentences. It can also be used to make explicit the set of conditions under which a non-tautology – i.e. a sentence which is a theorem of a particular theory $\mathscr{T}$[20] – is valid. The role played by the incorrect instances of ✠ rule is to identify the particular class of models validating the non-tautology into question. Differently from the case of tautologies or antilogies, where either the whole class or the empty class of models is considered, in the case of non-tautologies the instances of the ✠ rule single out a very specific and non-trivial class of models. It seems then that the proof-search dynamics leads to corroborate the standard view presented in Sect. 5.1 according to which the role of proper axioms is to identify classes of relational structures. However, the conceptual order is inverted here: structures are not primitive entities that are later syntactically fixed by the axioms, but they become generated from the syntactical features of the proof-search dynamics.

Nevertheless, these structures are not yet homogeneous with syntactical entities, since they are still set-theoretical entities. In what follows, we will present a general framework that allows to treat proofs and models – or better, countermodels – from a homogeneous point of view. In order to do that, we need to relax the usual notion of syntax. In order to make this idea clear, we will start with the problem of distinguishing between antilogies and non-tautologies.

5.4 Distinction Between Non-tautologies and Antilogies

The framework of $\mathrm{LK}_R^{✠}$ presented above does not yet allow to distinguish between sequents that are non-tautologies, and antilogies. Let us consider the two following derivations:

$$\dfrac{\dfrac{\dfrac{\dfrac{}{\vdash \neg A, B}\,✠}{\vdash \neg A \vee B}\,\vee \qquad \dfrac{}{\vdash A}\,✠}{\vdash (\neg A \vee B) \wedge A}\,\wedge \qquad \dfrac{\dfrac{}{\vdash \neg B}\,✠ \qquad \dfrac{}{\vdash \neg C}\,✠}{\vdash \neg B \wedge \neg C}\,\wedge}{\vdash ((\neg A \vee B) \wedge A) \wedge (\neg B \wedge \neg C)}\,\wedge$$

[20] In the literature, this kind of formulas are usually called *neutral formulas*. However, we prefer to avoid this terminology; even if these formulas are neutral form a logical point of view – they are neither tautology nor antilogy –, they are not neutral from the point of view of a particular theory T. Since our aim is to provide a framework that is applicable to specific mathematical theories, using this terminology could be misleading.

$$\dfrac{\dfrac{\dfrac{\dfrac{}{\vdash A, \neg B}\,\maltese}{\vdash A \vee \neg B}\,\vee \qquad \dfrac{}{\vdash A}\,\maltese}{\vdash (A \vee \neg B) \wedge A}\,\wedge \qquad \dfrac{\dfrac{}{\vdash \neg B}\,\maltese \qquad \dfrac{}{\vdash \neg C}\,\maltese}{\vdash \neg B \wedge \neg C}\,\wedge}{\vdash ((A \vee \neg B) \wedge A) \wedge (\neg B \wedge \neg C)}\,\wedge$$

The simple observation that the two proofs appeal to non correct instances of the ✠ rule does not tell us anything about the fact that the concluding sequent is an antilogy or a non-tautology. The only way to distinguish between antilogies and non-tautologies is to check if there exists a valuation rendering all ✠ rules true. For instance, such a valuation exists in the case of the second proof presented above (making A true, and B and C false), while it does not exist for the first one. We can thus conclude that the latter is a non-tautology, and the former is an antilogy. The problem with this way of distinguishing between non-tautologies and antilogies is that it appeals to the inspection of all possible valuations of all the ✠ rules in the proof. Hence, this method is not exclusively based on proofs' inspection; moreover, it is not really effective (especially when dealing with first order logic). Ideally, we want to be able to recognize a non-tautology or an antilogy by means of a simple mechanical inspection of the proof.

A possible way to decide if a sentence is a non-tautology or an antilogy is to analyze not only the proof of this sentence, but also the proof of its negation. First, it is worth recalling that we are working in a framework where everything is derivable: no particular constraints were imposed on the application of the $\maltese^{At}$ rule. For example, the $\maltese^{At}$ rule can be applied also when $\Gamma = \emptyset$, and thus the empty sequent "$\vdash$" can be derived in $\text{pLK}^{\maltese}_{R}$. In the object language, the empty sequent represents the idea that an unspecified absurdity[21] – namely, the empty succedent – is derivable from any kind of hypothesis – namely, the empty antecedent (see Paoli 2002, p. 32). Deriving the empty sequent corresponds therefore to deriving an absurdity as a theorem, and thus to showing that $\text{pLK}^{\maltese}_{R}$ is inconsistent. In order to prevent the system from being inconsistent, an ad hoc solution is to impose a constraint on the application of the $\maltese^{At}$ rule, namely that $\Gamma \neq \emptyset$. In this way, even if any formula can still become a theorem, the system remains consistent. We would be then in a situation complementary to the one advocated by paraconsistent logics: $\text{pLK}^{\maltese}_{R}$ would be a trivial but consistent system. In fact, it would be possible to obtain an empty sequent only by appealing to the cut rule, which would allow us to derive the empty sequent from the derivable sequents $\vdash A$ and $\vdash \neg A$, for a certain A. If it was the case, we would be able to characterize absurdity negatively, as something for which we do not possess a canonical derivation – that is a derivation terminating with the rule corresponding to the principal connective of the conclusion-formula. This is exactly the explanation of absurdity given by verificationist accounts (see Sundholm 1983, p. 485; Martin-Löf 1996, p. 51).

[21] Absurdity is seen here in a Brouwerian perspective, that is, as something that interrupts a derivation (Brouwer 1908, p. 109). In absence of any formula, no rule corresponding to a logical connective can be applied, and thus the derivation cannot be further carried on.

Such an explanation, however, does not fit with the notion of proof as characterized by the system $\mathrm{pLK}^{\maltese}_{R}$ (and more generally, $\mathrm{LK}^{\maltese}_{R}$). Firstly, using a multi-succedent calculus makes it difficult to identify the conclusion-formula of a derivation. Secondly, and more importantly, it is not always clear *what kind* of absurdity is obtained by cutting a proof of $\vdash A$ with one of $\vdash \neg A$. More specifically, if A is a tautology, then $\neg A$ is an antilogy, and therefore the empty sequent is also an antilogy. But if both A and $\neg A$ are non-tautologies, then we are not entitled to conclude that the empty sequent is an antilogy. We would then be in a situation where we have different proofs of the empty sequent but we cannot establish whether they are a proof of the same theorem. Despite such difficulties, we could still recognize different proofs of the same theorem if we had a cut elimination procedure that allows us to show that all these proofs of the empty sequent are reducible to the same cut-free proof. Now, cut admissibility for the system LK_{R} is a corollary of Schütte's completeness proof but this result is not effective with respect to proofs transformations.[22] The result simply states that if we have a proof with cuts then there exists a proof of the same sequent without cuts, but it does not provide an algorithm for transforming the proof with cuts into the cut free one, nor it tells us anything about the form of the cut free proof. As a consequence, it cannot then be used to establish the identity of proof results. If we want to define a cut elimination algorithm for $\mathrm{LK}^{\maltese}_{R}$ we will have to appeal to the admissibility of the structural rules of weakening and contraction. Here, the problem is that the cut elimination procedure defined in this way is not local, and it does not give any information about the evolution of the set of instances of the $\maltese$ rule during the process of cut reduction. This last point is particularly crucial because the system $\mathrm{LK}^{\maltese}_{R}$ is intended to be a tool for the study of what set of instances of the $\maltese$ rule are used in order to derive a particular sequent. If the set of instances changes during the process of cut elimination, then our analysis is likely to fail.

5.5 Liberalizing Syntax

In this section we define a framework that – like $\mathrm{LK}^{\maltese}_{R}$ – allows us to define logically incorrect proofs, but that at the same time also allows algorithmic cut elimination procedures that do not alter the set of proper axioms, i.e. the set of instances of the $\maltese$ rule. Our general aim is to produce a framework which is capable

[22] It would be incorrect to claim that Schütte's demonstration of cut admissibility through completeness is *tout court* non-effective. In a purely logical setting, given a valid sequent $\vdash \Gamma$ – i.e. $\vDash \Gamma$ – it is possible to effectively enumerate all the proofs of the theorems of LK_{R} until a proof of Γ is found. Since LK_{R} does not contain the cut rule, this is a cut-free proof. However, as Kreisel (1958, p. 167) remarks, « [it] is not an algorithm at all in the sense of a working mathematician, because it depends on 'trying out all proofs of the subject', i.e. it is a systematic method of trial and error. » In other words, the algorithm in question in not an *efficient* one.

of both (i) generating a semantics from syntactical procedures, and (ii) assigning an interesting computational interpretation to these procedures, an interpretation that is not limited to the mere availability of a proof search algorithm. In such a framework, the semantical role of axioms should be explained without appealing to the set-theoretic notion of a model, which is based on a primitive and epistemic-transcending notion of truth. In order to do that, we will define the notion of truth over that of proof.

This approach differs from the standard inferentialist one in that proofs are analyzed with respect to their computational content, while standard inferentialism only focuses on the order of applications of the rules. Within this perspective, proofs are not regarded as singular objects of study but they are always considered in connection with a given environment: if proofs correspond to programs, then their computational behavior can be detected only inside a context of evaluation, namely a context composed by other proofs/programs. Since proofs are studied in their interaction with other proofs by means of the Cut rule, the perspective adopted here can be characterised as a *global* perspective on proofs. On the contrary, a *local* perspective on proofs consists in studying how the structure of a given (single) proof can be rearranged by means of proofs transformations (e.g. rules commutations).[23]

In order to define a framework for a global account of proof, we need to liberalize our syntax. This is due to the fact that proofs are usually presented as trees, and this presentation forces us to interpret them as ordered sequences of inference rules. In contrast to this approach, we propose to look at proofs from a geometrical point of view, where the order of application of the rules is irrelevant. On this view, the emphasis is put on the "spatial" configuration of the premisses and conclusions of the rules, and on the transformations that can be operated on these configurations while keeping them invariant. Hence, the most appropriate objects for codifying such perspective are no longer the syntactical objects inductively generated by a grammar, but should be some kind of mathematical objects which do not necessarily represent ordered or inductive structures. Furthermore, the operations definable over these objects will have a computational content, so that the desired proofs-as-programs paradigm is respected.

Let us now present in some more detail the perspective just sketched.

[23]The local/global distinction is inspired by the terminology adopted in computer science to specify how the behavior of programs is studied. The application of this distinction has been first introduced by Paoli (2002) in order to analyze the inferential properties of proofs, and successively used by Poggiolesi (2011) and Hjortland (2012). More precisely, they claim that the meaning of logical constants depends both on the shape of the premisses and conclusions of the rules governing the inferential behavior of each specific connective, and on the way these rules interact with the others, particularly during the process of cut elimination.

5.5.1 *Proof Nets and Axioms*

The most suitable framework for liberalizing syntax is, in our opinion, *linear logic* (Girard 1987b). First of all, it should be noticed that adopting such a point of view does not put into question the classical point of view that we defend in this paper. The reason is that linear logic is nothing but a way to analyze classical logic at the microscope, namely by controlling the use of structural rules of weakening and contraction. This control is obtained from the decomposition of standard implication $A \to B$ into two distinct operations: a linear implication $\multimap$, and an exponential modality ! allowing the repeated use of the argument of type A. This refinement let emerge from the set of rules for classical logic two sets of rules: the set of rules with shared derivational contexts – also known as *additive rules* – and the set of rules with independent derivational contexts – also known as *multiplicative rules* (see Di Cosmo and Miller 2010, §2.1).[24] For the purposes of the paper we can limit our analysis to the multiplicative fragment of linear logic, **MLL**, which is composed of: a closure operator $\perp$ corresponding to an involutive negation (more on this will be said later), multiplicative conjunction $\otimes$, and its De Morgan's dual, i.e. multiplicative disjunction $\parr$. Linear implication, instead, is definable in the same way as in classical logic, i.e. $A \multimap B \equiv A^{\perp} \parr B$. The rules corresponding to these connectives are the following:

$$\frac{}{\vdash P, P^{\perp}}\ \text{ax}$$

$$\frac{\vdash \Gamma, A \quad \vdash \Delta, B}{\vdash \Gamma, \Delta, A \otimes B}\ \otimes \qquad \frac{\vdash \Gamma, A, B}{\vdash \Gamma, A \parr B}\ \parr$$

and the cut rule is:

$$\frac{\vdash \Gamma, A \quad \vdash \Delta, A^{\perp}}{\vdash \Gamma, \Delta}\ \text{Cut}$$

What makes linear logic particularly interesting for our discussion is the way in which proofs can be represented, or better, interpreted. The idea is to consider semantical entities which allow to deal with proof systems analogous to multi-conclusion natural deduction. In order to do that, we have to abandon the syntactic and linguistic analysis of proofs and replace it with a purely geometrical analysis. In this context, formulas are no longer interpretable as linguistic acts, but objets

[24] This is correct only in the case of binary rules, while it is less clear in the case of unary rules. When there is only one premiss there is only one context of derivation, and thus the problem of sharing it or splitting it does not arise. On the other hand, the presence of all the immediate subformulas of the conclusion of a unary rule in the premiss signals that in order to reconstruct the proof we need to follow all these formulas, since they are not derivable form the very same context. And this means that we are in presence of a multiplicative rule. Hence, the distinction we traced is not ambiguous as it could have been thought at first sight.

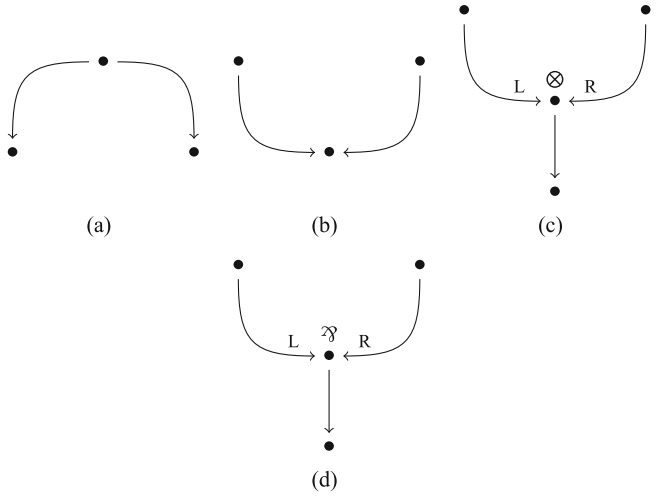


Fig. 5.4 Basic bricks for proof structures. (**a**) Axiom brick. (**b**) Cut brick. (**c**) Tensor brick. (**d**) Parr brick

organized according to certain spatial relations over which invariant transformations can be executed. In other words, a formula is identified by the position it occupies, and not by its syntactical form.

The notion of *proof nets* introduced by Girard (1987b) rests on this very idea. More precisely, a proof is represented by a graph[25] constructed from basic elements representing the axioms, the connectives, and the cut rule (see Fig. 5.4). A graph obtained in this way is called a *proof structure*.[26] Every sequent calculus proof can be represented as a proof structure, even though this correspondence is not injective. The non injectivity of this representation is the main motivation for the definition of proof structures which are meant to represent the quotient of sequent calculus proofs up to uninformative commutations of rules.

The main interest of proof nets lies in the fact that the syntax of proof structures is extremely tolerant, and it allows to construct graphs that do not come from a sequent calculus proof, such as the proof structures in Fig. 5.5. Those proof structures arising as representations of sequence calculus proofs are called sequentializable. Of course, this would not be helpful at all if we were not able to distinguish sequentializable proof structures: for this purpose one defines the notion of *proof net* as a proof structure that satisfies a given geometrical or topological property – called a correctness criterion, and then shows that proof nets are exactly the sequentializable proof structures (Fig. 5.6).

[25]Here we consider *directed simple graphs*, i.e. directed graphs with, for all vertices a, b, at most one edge of source a and target b.

[26]The terminological choice adopted here is intended to convey the idea that a proof refers to a structure not only as a syntactical entity, but also as a semantical one, as we have seen in Sect. 5.3.3 and will clarify later, especially in Sect. 5.5.2.

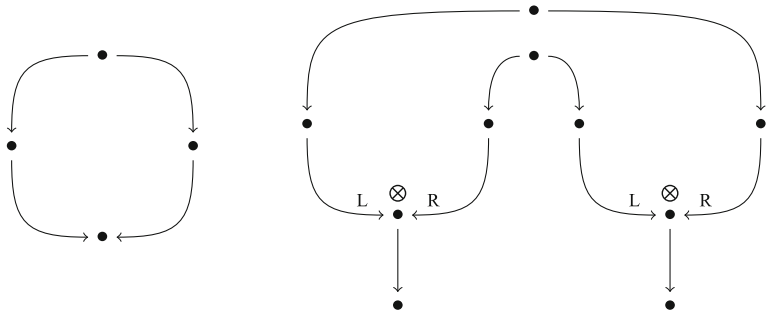
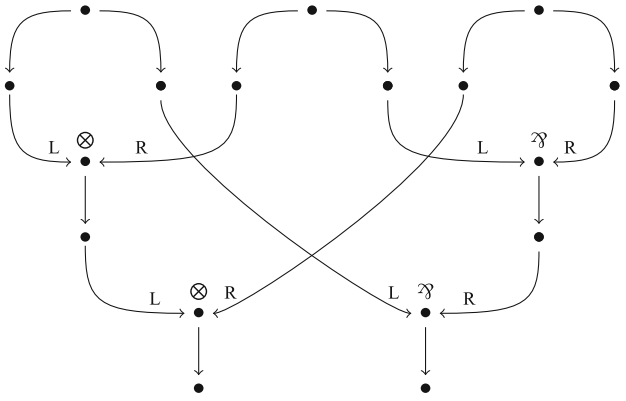


Fig. 5.5 Proof structures that are not proof nets

$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{}{\vdash P, P^{\perp}}\,\mathrm{ax} \quad \dfrac{}{\vdash Q, Q^{\perp}}\,\mathrm{ax}}{\vdash P \otimes Q, P^{\perp}, Q^{\perp}}\,\otimes \qquad \dfrac{}{\vdash R, R^{\perp}}\,\mathrm{ax}}{\vdash (P \otimes Q) \otimes R, P^{\perp}, Q^{\perp}, R^{\perp}}\,\otimes}{\vdash (P \otimes Q) \otimes R, P^{\perp}, Q^{\perp} \parr R^{\perp}}\,\parr}{\vdash (P \otimes Q) \otimes R, P^{\perp} \parr (Q^{\perp} \parr R^{\perp})}\,\parr$$

(a)



(b)

Fig. 5.6 Proof of associativity of the tensor: (**a**) Sequent calculus proof. (**b**) Proof net. Note that atoms have disappeared when representing the proof as a proof net. This illustrates the useful and deep peculiarity of proof nets: the same proof net represent the proof of associativity for the tensor between atoms P, Q, R as shown above, as well as the proofs of associativity for the tensor of any triple A, B, C of formulas. In a way, a proof net represents a scheme of proof more than a single proof

5.5.1.1 Correctness Criterions

Many correctness criterions are available (see Seiller 2012c, §2.3 for a survey), though they all share the same global idea (Seiller 2012c, pp. 24–25). Given a proof structure $\mathscr{R}$:

1. we define a family S of objects (call them *tests*);
2. we show that $\mathscr{R}$ is sequentializable if and only if all elements in S satisfy a given property.

The similarity runs deeper if we are a little more precise: in each case, the elements of S can be defined by a proof structure without its axioms, (i.e. a proof structure $\mathscr{R}$ where the axiom vertices and their ingoing/outgoing edges have been erased). The second part of the criterion then describes how the axioms interact with this axiom-less part of the proof structure, which will be denoted by $\mathscr{R}_t$. The only difference between two proof nets corresponding to the same formula A consists in their axioms, the graph $\mathscr{R}_t$ being defined uniquely from A. A set of axioms can thus be considered as an *untyped proof* – noted with $\mathscr{R}_a$ – and $\mathscr{R}_t$ as a type. The correctness criterion is then simply a typing criterion: if a set of axioms (an untyped proof) together with a type $\mathscr{R}_t$ yields a proof net, then this proof can be typed by the formula defining $\mathscr{R}_t$.

From now on, we will consider the correctness criterion based on the use of permutations (Girard 2011; Seiller 2012c)[27]:

- from $\mathscr{R}_a$ one defines a permutation σ_a;
- from $\mathscr{R}_t$, the set of tests S is defined as a set of permutations;
- $\mathscr{R}$ is a proof net if and only if for all $\tau \in S$, $\sigma_a\tau$ is a cyclic permutation.

We will say that two permutations σ, τ are *orthogonal* when their composition $\sigma\tau$ is cyclic.

Interestingly, the tests associated to $\mathscr{R}_t$ can be understood as counter-models. Indeed, if an untyped proof $\mathscr{R}_a$ cannot be typed by $\mathscr{R}_t$, it means that $\mathscr{R}_a$ is not a proof of the formula corresponding to $\mathscr{R}_t$. But the fact that $\mathscr{R}_a$ cannot be typed by $\mathscr{R}_t$ amounts to the existence of a test in S such that the product of σ_a (the permutation associated to $\mathscr{R}_a$) and τ is not cyclic. Showing that an untyped proof $\mathscr{R}_a$ is not a proof of a formula A then boils down to finding a test of A that is not passed by $\mathscr{R}_a$, in the same way in which a derivation in $\mathrm{LK}^{\maltese}_R$ can be shown incorrect by finding a counter-model that falsifies the set of axioms.

[27] Roughly speaking, the idea is to take a graph, drop all the formulas labelling its nodes, and label again only the nodes of the axiom bricks with natural numbers (counting from left to right). The different paths that can be defined through the graph – or through its subgraphs $\mathscr{R}_a$ and $\mathscr{R}_t$ – induce a set of permutations on the given (finite) set of natural numbers.

5.5.1.2 Cut Elimination

A cut elimination procedure can be defined; such procedure is compatible with the interpretation of proof nets $\mathscr{R}$ as a pair consisting in an untyped proof $\mathscr{R}_a$ together with a type $\mathscr{R}_t$. This cut elimination procedure is strongly normalizing and we can therefore choose particular strategies of reduction. Let $(\mathscr{R}_a, \mathscr{R}_t)$ and $(\mathscr{P}_a, \mathscr{P}_t)$ be two proof nets linked by a cut rule. We now consider the reduction strategies that first eliminate all cuts between $\mathscr{R}_t$ and $\mathscr{P}_t$, and then eliminate cuts between $\mathscr{R}_a$ and $\mathscr{P}_a$. This decomposition leads to the following interpretation:

- the cut elimination between types ensures that the specifications are compatible: if a cut cannot be eliminated then the strategy stops, which indicates that the two untyped proofs were not typed properly;
- the cut elimination between types, when successful, has no real computational meaning: it only defines a type $\mathscr{Q}_t$ and describes how the untyped proofs $\mathscr{R}_a$ and $\mathscr{P}_a$ are *plugged* together;
- the cut elimination between untyped proofs bears the computational content, and yields an untyped proof $\mathscr{Q}_a$ such that $(\mathscr{Q}_a, \mathscr{Q}_t)$ is a proof net.

5.5.1.3 Generalized Axioms

As it is the case for the framework described in Sect. 5.3, it is possible to extend the proof structure syntax by considering generalized axioms (Fig. 5.7).

In this setting, the generalized axioms represent a cyclic permutation and allow the derivation of any formula, in the same way as the as the ✠ rule allowed the derivation of any formula in the system $\mathrm{pLK}^{✠}_{R}$. But the change of paradigm, from sequent calculus to proof nets, reinforced the role of these generalized axioms. Once again, we can write such a generalized proof net as a pair $(\mathscr{R}_a, \mathscr{R}_t)$ composed by a type $\mathscr{R}_t$ and a *paraproof* $\mathscr{R}_a$. Paraproofs do not necessarily contain correct instantiations of axioms and rules. In this setting, one can prove that there is a correspondance between paraproofs of a formula A and the tests of its dual $A^{\perp}$. Let $\mathfrak{P}(A)$ denote the set of paraproofs $\mathscr{R}_a$ such that the pair $(\mathscr{R}_a, \mathscr{R}_t)$ is a proof net, where $\mathscr{R}_t$ is the type corresponding to a formula A. Then $\mathfrak{P}(A)$ corresponds to the orthogonal closure of the set of tests defined by the type $\mathscr{P}_t$ corresponding to the formula $A^{\perp}$.

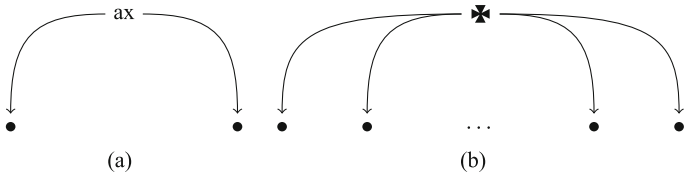


Fig. 5.7 Generalizing axioms in proof structures. (**a**) Identity axiom. (**b**) Generalized axiom

The circle is now complete:

- an untyped paraproof can be given a type A if and only if it is orthogonal to the tests for A;
- an untyped paraproof is a test for the type A if and only if it is orthogonal to the proofs of A, i.e. proofs are tests for tests.

These remarks on proof structures support the idea that generalized axioms are a way of adding counter-models to the syntax. As we will see in the next section, this idea can be even used to redefine a logic where the objects are generalized untyped proofs, and the formulas are defined interactively.

5.5.2 Untyped Proof Theory

The ideas expounded in Sect. 5.5.1.3 lead to the definition of the first version of a *geometry of interaction*, where basic objects are permutations (Girard 1988). This construction was then generalized to include of more expressive fragments of linear logic. In this section, we will describe this type of constructions in a very general way. We will focus in particular on *Ludics* (Girard 2001), which inspired the ideas developed in Sect. 5.6. Such a framework will be called *untyped proof theory*.

Definition 5.29. An untyped proof theory is given by:

- A set of *untyped paraproofs* $\mathscr{U}$;
- A notion of execution $\mathscr{U} \times \mathscr{U} \to \mathscr{U}$, denoted here by $a, b \mapsto a :: b$;
- A notion of termination given by a set of untyped proofs $\Omega \subset \mathscr{U}$.

We can then construct everything from the notion of execution. First, the notion of *orthogonality* is defined: two paraproofs a, b are orthogonal – denoted $a \perp b$ – if and only if their execution $a :: b$ is in Ω. It is worth noting that no particular constraints are imposed on Ω; this means that the notion of termination is not absolute, but relative to what we consider a terminating configuration for the untyped proofs under analysis. Hence, untyped proof theory represents an extremely flexible computational framework. From a more mathematical point of view, the notion of orthogonality intuitively corresponds to the phenomenon that occurs between generalized axioms in proof structures: if $a \perp b$, then a (resp. b) must be a paraproof of a formula A (resp. $A^{\perp}$), or equivalently a test for $A^{\perp}$ (resp. for A). Following up on this parallel, *types* can be defined as the sets of paraproofs that are equal to their bi-orthogonal. Equivalently, a type is defined as a set of paraproofs A such that there exists another set of paraproofs B with $A = B^{\perp} = \{a \mid \forall b \in B, a \perp b\}$, i.e. a type is a set of paraproofs that pass a given set of tests B.

We can now define logical constants as constructions on paraproofs; these constructions in turn induce constructions on types. Specific constructions would

then allow to recover fragments or even full linear logic. Below are three examples of such ideas.[28]

Example 5.30. The first construction based on permutations mentioned at the beginning of this section can be enriched in order to obtain a construction for **MLL** with units (Seiller 2012b), for **MALL** with additive units (Seiller 2012a), for Elementary Linear Logic (Seiller 2013, 2014) – a subsystem of linear logic that characterizes the set of functions computable in elementary time. In this setting, the set of paraproofs is a set of pairs of a graph and a real number, the notion of execution is based on the graph of alternating paths between two graphs, and the notion of termination is given by the set of pairs $(a, \emptyset)$, where $a \neq 0$ is a real number and $\emptyset$ denotes the empty graph on an empty set of vertices.

Example 5.31. In Ludics, the set of untyped paraproofs is the set of *designs-desseins*, the execution is the cut elimination procedure over these objects, and the notion of termination is the set of dessein containing a single design: the daimon.

Example 5.32. In the latest version of geometry of interaction, the set of untyped paraproofs is defined as the set of *projects*, while the execution is defined as the solution to the *feedback equation* and the termination is defined as the set of conducts of empty carrier with a non-null wager.

It is the possible to characterize in theses frameworks which paraproofs correspond to proofs in the same way as it is possible to identify correct derivations among the derivations of $\mathrm{LK}_R^{\maltese}$. These objects, which are called *successful* – or *winning*, to emphasize their relation to winning strategies in game semantics – can be tested against unsuccessful ones. The latter correspond to counter-models. We are thus in a framework where the syntactical and semantical (in the classical sense) aspects of logic are both represented in a homogeneous way. More specifically, distinguishing antilogies from non-tautologies does not involve anymore the verification of each counter-model of some formula A, but only requires deciding whether the set of tests of A, i.e. $A^{\perp}$, contains a successful paraproof.

5.6 Philosophical Considerations

Hitherto we presented some ideas for developing a computational account of proofs which is general enough to allow the justification of logical statements, as well as of proper axioms. However, our general aim is to show that this computation setting can constitute an appropriate framework for the development of a uniform and epistemic-based understanding of logic and mathematics. In order to do this, we will adopt a linguistic point of view principally based on the analysis of the meaning

[28] A full analysis of logical constants from the untyped perspective is part of ongoing researches. See Naibo et al. (2011).

of logical and mathematical sentences. Our meaning-theoretical analysis will be focused on the untyped framework presented in Sect. 5.5, and based on Girard's geometry of interaction.[29]

5.6.1 Normative vs. Descriptive Theories of Meaning

As we said in Sect. 5.1, in this paper we made an attempt of reconciling a standard notion of axiom with a certain kind of inferentialism based on the computational interpretation of proofs, or better, of proof structures. However, we have not yet clarified the philosophical extent of this kind of inferentialism. In particular, we have not yet made explicit which kind of theory of meaning can be induced by this computational perspective. Our claim is that such a theory of meaning is rather different from the one associated to standard inferentialism; the main difference being that the latter is *normative* while the former is *descriptive*. Let us try to clarify this crucial point.

Standard inferentialism is usually identified with Dummett-Prawitz verificationism (cf. Tennant 2012, §2). On to this approach, the rules governing our linguistic practice – i.e. the rules we are supposed to master in order to have successful linguistic exchanges – have to be governed by a principle of harmony that prevents the generation of new informations in a non-conservative way. This principle can be formally captured by the *inversion principle* formulated by Prawitz (1965, p. 33; 1973, pp. 232–233): anything that follows from the assertion of a certain (complex) sentence A, cannot exceed what directly follows from the grounds for asserting it, i.e. from the premisses of the introduction rule for A. This principle plays the role of a *norm* because it transcends the situation that it regulates: by imposing it at the beginning of the construction of a language, it guarantees in principle a "perfect" communication, avoiding misunderstandings as well as other linguistically pernicious situations (cf. Dummett 1973a, p. 454, for the well known example of 'Boche').[30] It would not be an exaggeration to say that from the verificationist point of view what counts is the way in which the speakers structure the contents of what that they want to communicate (i.e. the messages). If these contents are structured using expressions which respect the inversion principle, then this would already

[29] A similar analysis is proposed by Bonnay (2007) on the basis of Krivine's classical realizability (see Krivine 2003).

[30] This normative conception of language shares some similarities with the *Orthosprache-project* promoted by the so-called *Erlangen School* (or *Erlangen Constructivism*; Kamlah and Lorenzen 1972; Lorenzen and Schwemmer 1973). On this view, « language is not just a fact that we discover, but a human cultural accomplishment whose construction reason can and should control » (Rahman and Clerbout 2013, p. 4). However, the *Orthosprache-project* differs from Dummett-Prawitz verificationism in that it endorses a pluralistic – and not a monistic – conception of logic which allows to justify, for example, both intuitionistic and classical logic (see Rahman and Clerbout (2013), p. 9, 67, note 28; Sørensen and Urzyczyn (2006), §§4.5,4.6, 6.5,7.5).

be sufficient in order to guarantee communication to work correctly. From such a perspective the responsibility for a good communication completely rests on the person who sent the message and not on the one who receive it.

Both the computational perspective we adopted in this paper and standard inferentialism take proofs to be the meaning-conferring objects. However, there is an essential difference between the two perspectives. As already seen in Sect. 5.5.1.3, the computational perspective asks programs (i.e. paraproofs) to be tested in order to evaluate their behavior, and testing requires a context of evaluation (i.e. a set of paraproofs). Under the proof-as-programs correspondence, proofs are still necessary to determine the meaning of a sentence, but they are no longer sufficient. More specifically, knowing the order in which the rules have been applied in a proof is not sufficient: we also need to know what play the role of context of evaluation. In the perspective presented here, this is achieved by using the notion of paraproof. As we have seen, paraproofs do not necessarily represent correct – i.e. logically valid – (linguistic) arguments: the correctness of a paraproof depends on the interactional properties it displays in the presence of other paraproofs. From the linguistic point of view, if a proof corresponds to a correct justification for the *assertion* of a sentence (i.e. a correct justification for judging that sentence as true; cf. Martin-Löf 1987; Sundholm 1997), a paraproof corresponds to an *argument* supporting the *utterance* (see Lecomte and Quatrini 2011a) of a certain sentence in a particular context of discourse, regardless of whether the argument is (logically) correct and the sentence is true. In the same vein, the process of interaction between two paraproofs can be seen as a dispute between two speakers who use arguments to convince each other to accept their own position. In this sense, truth ceases to be an absolute notion and becomes an interactional and "social" one: a sentence can be judged as true when the speaker *always* possesses a convincing argument, i.e. when the speaker possesses a winning strategy (cf. p. 165, *supra*).

A further fundamental feature of this framework is that the meaning of sentences is not fixed by a set of rules obeying pre-established principles, but it is determined within the linguistic activity itself: knowing the meaning of a sentence corresponds to knowing which counter-arguments can be used in a dispute in order to terminate it. Since these counter-arguments strictly depend on the specific context and situation considered – namely, the arguments used by the other speaker – this means that they cannot be determined in advance of a dispute nor from "outside" the linguistic exchange itself. In this sense, the untyped approach induces a sort of game-theoretic semantics.

An important difference with standard game-theoretic semantics[31] is that in this framework knowledge of the meaning of a sentence does not correspond to

[31]Classic texts in game-theoretic semantics are Hintikka (1983) for a model-based account of games, and Lorenzen and Lorenz (1978) for a syntactical and operative – or dialogical – account. For detailed surveys of recent developments of dialogical approaches to game-theoretic semantics see Rahman and Keiff (2004) and Keiff (2009). For textbook presentations see Redmond and Fontaine (2011) and Rückert (2011).

knowledge of how to win a dispute. In order to determine the meaning of a sentence it is only requested that the dispute terminate; whether it does so with a gain or with a loss is irrelevant. As we mentioned before, having a winning strategy corresponds to knowing that the sentence in question is true. This implies that the meaning of a sentence neither coincides with a truth-definition nor depends on a primitive, unanalyzed notion of truth. It is for this very reason that, in Dummettian terms, the computational and untyped approach to semantics can be characterized as an anti-realist position: « [...] the notion of *truth*, considered as a feature, which each mathematical statement either determinately possesses or determinately lacks, [...] cannot be the central notion for a theory of the meanings of mathematical statements », on the contrary, « [...] it is in the mastery of [a] practice that our grasp of the meaning of the statements must consists » (Dummett 1973b, p. 225). Furthermore, characterizing the truth of a sentence as the possession of a winning strategy also allows to do justice to Dummett's *manifestability* requirement (Dummett 1973b, pp. 93–95; Dummett 1976, pp. 79–82; Dummett 1977, pp. 193–195): the fact that a sentence is true makes a noticeable difference at the level of our linguistic practice, since it implies that there is someone who would always gain a dispute if they were to argue for it.[32]

To sum up, the computation-based theory of meaning introduced here is still within the "meaning as use" paradigm. A distinguishing feature of this theory with respect to other "meaning as use" theories is that the set of licensed uses of a sentence is not defined by an absolute and external norm, but by the dispositions of the other speakers to respond to those uses. In this sense the notion of the correct use of a sentence emerges from the linguistic practice itself. More generally, standard inferentialist theories of meaning hold that the aim of a theory of meaning is to fix *in abstracto* the rules that a language must respect in order to work properly, i.e. in order to do what we expect it to do – allow communication between speakers. The perspective adopted here departs from standard inferentialism on this respect. In analogy with the position endorsed by Wittgenstein in the *Philosophical Investigations*, we hold that the aim of a theory of meaning is not to determine the « essence of a language » (Wittgenstein 1956, §97), that is, the set of characteristic features that a (abstract and idealized) language should possess. The mere fact that there is an established linguistic activity[33] and that it allows successful communication should lead us to think that linguistic activity should be considered as something that already works properly, not as something that should be rectified (Wittgenstein 1953, §98). From this perspective, linguistic ambiguities, usually considered to be sources of possible misunderstandings, are considered to be proper parts of the linguistic activity and as such, they are not rejected as incorrect.[34] This is a natural consequence of the absence of any a priori principles distinguishing

[32] A similar idea is presented in Marion (2012).

[33] This fact can be established "empirically".

[34] Indeed, in Ludics it is possible to represent fallacies in a formal and precise way as it has been shown by Lecomte and Quatrini (2011b).

between correct and incorrect (instances of) sentences. In other words, the idea is that the rules governing our linguistic activity are *immanent* to it. This view has a twofold consequence. On the one hand, we become aware of the way in which meaning is assigned to sentences by describing the linguistic activity. On the other hand, the knowledge of the meaning of a sentence is manifested in the capacity of taking part in a linguistic exchange where this sentence is used. On this view there is no need to make the rules governing the linguistic exchange explicit – if this where so, we would be forced to adopt an externalist approach. This feature calls attention to a first difference with standard Dummett-Prawitz verificationism. Other relevant differences will be analyzed in the following sections.

5.6.2 Feasibility and Interaction

The computational perspective described above is compatible with an inferentialist point of view as long as the application of an inference rule within a paraproof corresponds to the successful performance of a linguistic act within a linguistic exchange. The peculiarity of our perspective consists in the fact that the choice of which rule to apply is constrained by the type of linguistic acts previously performed both by the speaker and by her opponent. In other words, the speaker chooses what rules to apply strictly on the basis of the specific linguistic situation she is confronted with. This has two main consequences. First, inference rules act on linguistic objects that – as utterances – are situation-dependent; they do not act on more abstract or "absolute" linguistic entities such as assertions (see Lecomte and Quatrini 2011a).[35] Second, the fact that linguistic acts are situation-dependent means that the inference rules used to perform them have to take into consideration the particular type of resources available in each situation.[36]

In this respect, the theory of meaning which emerges from the computational untyped setting presented above is still characterized by anti-realist features, as it is articulated on the basis of the linguistic competences possessed and manifested by the speaker, rather than on a primitive and unanalyzed notion of truth. Nonetheless, it retains certain differences with respect to standard forms of anti-realism, such as Dummettian verificationism. More specifically, in our computational untyped setting mastering the meaning of a sentence does not consists in knowing what

[35]Furthermore, paraproofs are not necessarily correct proofs and thus their definition does not involve the notion of truth. On the contrary, in the Bolzanian tradition an assertion takes the form of the judgment 'A is true' (where *A* is a sentence or a proposition). Thus, both in the case in which assertion is taken to be a primitive notion – e.g. by realist positions – and in the case in which is taken to be non primitive – by anti-realist positions – assertion is defined is defined with respect to the very notion of truth.

[36]These considerations become evident when we consider that the untyped setting introduced here validates the logical rules of linear logic; it indeed widely acknowledged that linear logic is the clearest example of a resources-oriented logic (Di Cosmo and Miller 2010).

can be done with that sentence *in principle*, but it consists in knowing what can be *practically* done with it in some particular situations. In this sense, accepting a computational untyped approach leads to accepting some sort of *radical anti-realist* position: to know the meaning of a sentence is to know how it can be *feasibly* used during a concrete linguistic exchange. The computational approach does not consider idealized scenarios, but it focuses on concrete dialogical situations: its aim is not to determine the principles necessary for the construction of a language, but to *represent* (the abstract structure of) a language. For this reason, practical constraints are not obtained by imposing on verificationist's principles a constraint on proof-size bounds in advance,[37] for example by imposing a polynomial growth of proof-length by changing the usual connectives with linear ones during normalization or cut elimination (see Dubucs 2002; Dubucs and Marion 2003)[38]: such a change of connectives could only be justified ad hoc. On the contrary, it is the very nature of the interactional approach that ensures the existence of bounds which guarantee that the knowledge of the meaning of sentences is based on linguistic skills manifestable by the speakers: the presence of two speakers, instead of only one, guarantees that the actions of each speaker are always constrained by the actions of the other one, and vice versa.

[37]These kinds of bounds are essentially dictated by two reasons: (1) guaranteeing that the process of *verification* that something is a proof can be practically done by human beings; (2) guaranteeing the semantic key objects, i.e. canonical proofs, to be objects that can be practically *constructed* by human beings. By respecting these two conditions it should be assured that, in the verificationist account, both truth and meaning never make appeal to entities transcending concrete human capacities, as it could be the existence of proofs the size of which goes beyond physical limits.

[38]The standard justification for the choice of polynomial bounds can be found in Wang (1981, §6.5) and it has been well summarized by Marion (2009), p. 424:

> It is generally agreed that polynomial-time computability captures the capacities of digital computing machines, as opposed to their *idealized* counterparts, the Turing machines. Digital computing machines do *not* have access to unlimited resources, and this seems to be the key point for a radical anti-realist program. It is only asked here from the radical anti-realist that she grants that digital computing machines are an unproblematic extension of human cognitive capacities, so that, with polynomial-time computability, one remains within the sphere of what is humanly feasible.

In fact, it seems to us that there is a further, and usually neglected, argument supporting this choice. Schematically, it can be presented in the following way: (i) from the verificationist point of view meaning is based on proofs, and the only logical rules allowed for constructing these proofs are the intuitionsitic ones; (ii) via the Curry-Howard correspondence each proof of intuitionistic logic can be associated to a computable function, and *vice versa*; (iii) a fundamental property of a theory of meaning is compositionality; (iv) by restricting to polynomial computable functions, compositionality between functions (i.e. proofs) is preserved: the composition of two polynomial computable functions is still a polynomial computable function.

5.6.3 Holism and Molecularism

Our last observations will concern a typically Dummettian theme, namely the debate between molecularist accounts of meaning and holistic ones.

Generally speaking, it is usually argued that the adoption of an axiomatic approach comes with the acceptance of some kind of holism (see Troelstra and van Dalen 1988, pp. 851–852). This is because presenting a theory in an axiomatic way has two main consequences with respect to our understanding of the set of sentences constituting the theory itself. Firstly, the syntactical behavior of the expressions composing the language is fixed holistically: an expression is defined on the basis of the relations it entertains with the other expressions of the language, and there is no bound fixed in advance on the number of expressions that can be mutually related by the axioms. Secondly, the inferential behavior of an axiom can be fully determined only when it is used in conjunction with other formulas in order to extract some relevant information from it, and also in this case no bound can be imposed on the size of the set of these formulas in advance.[39] This picture seems to clash with the molecularist approach defended by Dummett. On such view, mastering a limited and well-determined fragment of a language is sufficient to understand the meaning of a given sentence – that is, linguistic competence does not require an ability to master the whole totality of the expressions of a language (Dummett 1976, p. 79). However, this aspect appears to be in conflict with the axiom-based perspective adopted in this paper. Thus, the question arises of whether our approach is *essentially* holistic or whether it could be compatible with a kind of molecularism akin to the Dummettian perspective. More specifically, while the untyped computational perspective allows to define types – and thus also formulas and sentences – as sets of paraproofs (cf. §5.2 *supra*), it does not provide a standard inductive definition of them. In particular, there is no such notion as that of atomic type.

Prima facie, this seems to contrast with Dummett molecularism in so far as this requires the possibility of ranking sentences in a hierarchy of increasing complexity.[40] In absence of a way of fixing such a hierarchy of types-formulas, we might be unable to impose a bound on the complexity of the set of formulas that have to know in order to understand another given formula. Now, as seen above, in a computational framework knowing the meaning of a sentence amounts to being able to participate in a linguistic exchange once the sentence in question occurs. In absence of a way of fixing in advance a hierarchy of types-formulas according to their complexity, it may seem that there is no way of fixing in advance the kind of formulas that will be involved in the exchange. It would then follow that the

[39]The other way round, this situation corresponds to the idea that to understand the meaning of an axiom it is necessary to understand the *totality* of the consequences that can be drawn from it (see Dummett 1991, p. 228).

[40]In particular, see Dummett (1991, p. 223): « Compositionality demands that the relation of dependence imposes upon the sentences of the language a hierarchical structure deviating only slightly from being a partial order. »

understanding of the sentence in question may only be explained by appealing to the understanding of all the expressions of the language, which would make our perspective incompatible with Dummett's analysis. In the remainder of this section, we will argue that our perspective is indeed compatible with Dummettian verificationism.

Dummettian verificationism holds that understanding the meaning of a complex sentence A is reducible to understanding the meaning of its principal connective; in order to do so, we only need to consider the set of sentences in which this connective appears as the principal connective. This means that only a limited fragment of the language has to be analyzed in order to understand the meaning of a certain expression. This analysis can be carried out by focusing on the properties of the inference rules involved in the (direct) justification of A. This amounts to the ability to recognize what counts as a canonical proof of A and whether the inference rules used in the (putative) proof are correct, i.e. valid. The correctness of the rules is usually ensured by the inversion principle we mentioned in Sect. 5.6.1: what can be drawn from the elimination rules of a certain connective must already be drawn from the premisses of its corresponding introduction rules. As Sundholm (2004, p. 454) remarked, the peculiarity of this principle is that it « [...] leads straightforwardly to a resurrection of the old idea that the validity of an inference resides in the analytic containment of the conclusion in the premisses ». It is this very possibility of reducing proofs to "analytic proofs" that plays a crucial role in the molecularist approach: from the syntactical form of a given complex sentence A it is possible to extract a relevant information which allows to impose a bound on the set of sentences necessary to understand A.[41] In particular, if the inversion principle is respected, it could be possible to prove the so-called *subformula property*, which guarantees that if a complex sentence A is provable, then there exists a proof the rules of which are applied only to subformulas of the conclusion.[42]

Despite the analogies between the Dummettian and the computational approaches, we can see more precisely that there are some crucial differences. In analogy with the Dummettian perspective, the computational approach reduces the understanding of the meaning of a sentence A to the understanding of the principal connective of A. This may lead to thinking that it is sufficient to consider only a

[41] It is worth noting that Girard's latest work, known under the name of *transcendental syntax*, aims at studying these issues from a formal point of view. In particular, it tries to prove how in some specific cases – namely when purely logical formulas are considered – the sets of tests can be shown to be *finite* (see Girard 2013, in part. § 3.2). The framework developed by Girard perfectly fits into our definition of untyped proof theory, as it is a particular case of Seiller's interaction graphs construction (Example 5.30).

[42] In fact, sometimes it could already be sufficient to prove a weaker property, like the subterm property. There are some theories – like the theory of equality, of groupoids and of lattices – for which the fact that a proof of A can make appeal to no other terms than those appearing in A is already sufficient to impose a bound on the set of formulas that should be known in order to know the meaning of A. The reason is that, from a technical point of view, for these theories the subterm property works like the subformula property: it allows to define proof-search methods by limiting the proof-search space (cf. Negri and von Plato 2011, §4).

fragment of the language in order to understand A, namely the set of sentences in which the principal connective of A is principal. However, unlike the verificationist approach, the computational perspective presented here does not assume that the introduction rules and the harmony requirement do, by themselves, confer meaning on a certain connective. This is because knowing the meaning of A – or better, the meaning of its principal connective – does not require knowing how to directly justify A, but instead requires knowing how to construct an argument against A. What counts is to have a strategy to refute A, or equivalently, a strategy in support of $\neg A$. Thus, the meaning-conferring objects are represented by argumentative strategies as a whole, and not by single inference rules. The only property that these strategies are required to have is to terminate when an argument in support of A is introduced. More specifically, there are no a priori contraints on the order of the inference rules in the argument for $\neg A$, and it is not required that all the rules applied in the argument are correct (i.e. valid). In summary, no analyticity constraints are imposed on these arguments. What really matters is the strategy that must be followed in order to refute A, while the formulas used in applying this strategy take a back seat. The upshot is that there are no a priori limitations on the formulas involved in the arguments used for refuting A, and this leads towards a holistic account of the computational approach presented in this article.

Lastly, we would like to consider an objection to this interpretation. It may be thought that the claim that our computational approach and Dummettian verificationism are compatible is in conflict with what we said in Sect. 5.6.2 about feasibility properties, and therefore with some kind of internal limitations that seem intrinsic to the computational approach. We argue, however, that the contradiction is only apparent. In Sect. 5.6.2, we saw that the knowledge of the meaning of a sentence by a certain speaker is manifested by participating in a linguistic exchange with another speaker, where this exchange involves only a bounded amount of resources. What is relevant here, instead, is the impossibility of establishing in advance which fragment of the language the speaker has to master in order to perform linguistic exchanges with other speakers. This impossibility is not a particularly surprising feature for a "computational theory of meaning".

As discussed above, the computational perspective comes with a certain understanding of what a *descriptive* theory of meaning is. From this perspective, what counts are the competences of those speakers who effectively participate in linguistic exchanges, but there is no requirement to fix those competences in advance. In particular, there is no need to explain what features a language must possess in order to be learnable. On the contrary, by focussing on the molecularity property we can see that the problem of *learning* a language plays a special role in the verificationist theory of meaning (see Dummett 1993, p. ix). Human agents can process only a limited amount of information at a time, but can nevertheless learn languages. Learning the meaning of an expression therefore requires mastery of only a finite fragment of the language in question (see Dummett 1973a, p. 515). Otherwise the task would go beyond human capacities, which are *ex hypothesi* finite ones.

5.7 Conclusion

The aim of this paper is to give a new account of the meaning of mathematical axioms that does not appeal either to a primitive notion of truth or to other realist assumptions. Our approach can therefore, broadly speaking, be characterized as an anti-realist one.

A major difficulty arises when one tries to interpret the notion of axiom through a Dummettian anti-realist semantics. The addition of axioms to standard proof systems entails the loss of (the notion of) canonical proofs, which is in fact the cornerstone of a verificationist theory of meaning. The existing solutions to this problem require a deep change in the epistemological status of axioms: axioms are turned into specific kind of rules. As a result, these solutions lead to revisionist positions with respect to the architecture of mathematical theories. Our computational approach overcomes these difficulties, at least in part, by enriching the set of primitive semantic concepts in the underlying theory of meaning. Moreover, our approach is compatible with an inferentialist, anti-realist understanding of meaning.

Our strategy in this paper was twofold. First, we explored the computational aspects of a proof, considered as an "isolated" object, *via* proof-search algorithmic techniques. A careful analysis of the occurrences of the ✠ rule allowed us to show a precise correspondence between (logically incorrect) generalized axiom rules and counter-models using a homogenous proof-theoretical setting. Secondly, we explored the computational aspects of the interaction between proofs, considered as objects interacting through the Cut rule. This approach allows one to forget formulas, by focusing only on the geometry of rules and their interactions. In this setting, generalized axioms provide a characterization of the crucial notion of paraproof. Both of these computational viewpoints reinforce the idea that generalized axioms are a way of working with (counter-)models inside the syntax. Axioms can thus be seen as fundamental entities at the crossroad between the syntax and the semantics of proof systems.

In the final part of the paper we made explicit some philosophical assumptions allowing us to integrate the analysis of untyped proofs with an inferentialist theory of meaning. We carried out such an analysis by pointing out some crucial differences between the inferentialist account based on Dummettian verificationism and the one based on the interactional approach presented in Sect. 5.5. Despite the fact that neither account considers the notion of truth as primitive but as epistemically dependent, a major divide exists between them. It amounts to the difference between a normative (and "solipsistic") theory of meaning and a descriptive (and "social") one. In the end, we showed in which sense the shift from the former to the latter leads us to embrace an even more radical form of anti-realism. Finally, we concluded our work with a short discussion around holism and molecularity. In order to fully comprehend the theory of meaning standing behind the computational and interactional approach presented in this paper, it is necessary to establish whether this theory of meaning is more harmonious with a holistic or a molecularistic

approach. The answer to this question is not yet established and seems to us a valuable direction for future research.

Acknowledgements We would like to thank Vito Michele Abrusci, Marianna Antonutti Marfori, Clément Aubert, Mathieu Marion, Shahid Rahman, and Luca Tranchini for their interest in our work, and for their useful comments and suggestions. We also wish to thank Myriam Quatrini, Jean-Baptiste Joinet, Damiano Mazza, Luiz Carlos Pereira, and Alain Lecomte for invitations to present this work in Marseille, Lyon, Paris, Rio de Janeiro, and Rome, respectively. Finally, we thank an anonymous referee for valuable suggestions that helped to improve the article.

This work has been partially funded by the French-German ANR-DFG project *Hypothetical reasoning: its proof-theoretic analysis – HYPOTHESES* (ANR-11-FRAL-0001) and by the French-German ANR-DFG project *Beyond Logic* (ANR-14-FRAL-0002).

Appendices

A Properties of System $\mathrm{pLK}^{✠}_{R}$

A.1 Proof of Proposition 5.3

By induction on the number of connectives in the sequent. Since the proofs for $\vee$ and for $\wedge$ are similar, we only show the invertibility of the $\vee$. The base case for the induction is a sequent containing only one connective, i.e. a ✠ rule followed by a $\vee$-rule. In this case the $✠^{At}$ rule itself gives a derivation of the premiss of the $\vee$-rule. Let us now assume that this is true for any sequent containing at most n connectives, and let us take a derivable sequent $\vdash \Gamma, A \vee B$, containing $n+1$ connectives, and let π be one of its derivations. If π ends with a $\vee$-rule, the subderivation obtained dropping the last rule gives a derivation of the premiss. If π does not end with a $\vee$ rule, then it must end with a $\wedge$ rule:

$$\dfrac{\begin{array}{cc}\pi_1 & \pi_2 \\ \vdots & \vdots \\ \vdash \Delta, A \vee B, C & \vdash \Delta, A \vee B, D\end{array}}{\vdash \Delta, A \vee B, C \wedge D}\ \wedge$$

Applying the induction hypothesis on the premisses we get two derivations π_1' and π_2' of $\vdash \Delta, A, B, C$ and $\vdash \Delta, A, B, D$ respectively; thus the desired derivation is:

$$\dfrac{\begin{array}{cc}\pi_1' & \pi_2' \\ \vdots & \vdots \\ \vdash \Delta, A, B, C & \vdash \Delta, A, B, D\end{array}}{\vdash \Delta, A, B, C \wedge D}\ \wedge$$

A.2 Proof of Lemma 5.6

By induction on the number of connectives in $\vdash \Gamma$. The base case is obvious. Assume that the lemma is true when the number of connectives in Γ is at most n. We show that the derivations of $\vdash \Gamma, F, G$ terminating with a rule where F is principal have the same set of leaves as the derivations terminating with a rule where G is principal. We only show how the proof is done in the case $F = A \wedge B$ and $G = C \wedge D$, which is the more complicated case. By the invertibility of the $\wedge$-rule we have the two derivations:

$$\dfrac{\dfrac{\dfrac{\begin{matrix}\pi_1 \\ \vdots \\ \vdash \Gamma, A, C\end{matrix} \quad \begin{matrix}\pi_2 \\ \vdots \\ \vdash \Gamma, A, D\end{matrix}}{\vdash \Gamma, A, C \wedge D}\wedge \quad \dfrac{\begin{matrix}\pi_3 \\ \vdots \\ \vdash \Gamma, B, C\end{matrix} \quad \begin{matrix}\pi_4 \\ \vdots \\ \vdash \Gamma, B, D\end{matrix}}{\vdash \Gamma, B, C \wedge D}\wedge}{\vdash \Gamma, A \wedge B, C \wedge D}\wedge$$

$$\dfrac{\dfrac{\dfrac{\begin{matrix}\rho_1 \\ \vdots \\ \vdash \Gamma, A, C\end{matrix} \quad \begin{matrix}\rho_2 \\ \vdots \\ \vdash \Gamma, B, C\end{matrix}}{\vdash \Gamma, A \wedge B, C}\wedge \quad \dfrac{\begin{matrix}\rho_3 \\ \vdots \\ \vdash \Gamma, A, D\end{matrix} \quad \begin{matrix}\rho_4 \\ \vdots \\ \vdash \Gamma, B, D\end{matrix}}{\vdash \Gamma, A \wedge B, D}\wedge}{\vdash \Gamma, A \wedge B, C \wedge D}\wedge$$

Using the induction hypothesis, we have that $\mathfrak{L}(\pi_k) = \mathfrak{L}(\rho_k)$ for all k in $\{1, 2, 3, 4\}$. Using these equalities and the induction hypothesis once again, we obtain that:

- any derivation π of $\vdash \Gamma, A, C \wedge D$ satisfies $\mathfrak{L}(\pi) = \mathfrak{L}(\pi_1) + \mathfrak{L}(\pi_2)$;
- any derivation π of $\vdash \Gamma, B, C \wedge D$ satisfies $\mathfrak{L}(\pi) = \mathfrak{L}(\pi_3) + \mathfrak{L}(\pi_4)$;
- any derivation π of $\vdash \Gamma, A \wedge B, C$ satisfies $\mathfrak{L}(\pi) = \mathfrak{L}(\pi_1) + \mathfrak{L}(\pi_3)$;
- any derivation π of $\vdash \Gamma, A \wedge B, D$ satisfies $\mathfrak{L}(\pi) = \mathfrak{L}(\pi_2) + \mathfrak{L}(\pi_4)$;

We have therefore shown that if π is any derivation of $\vdash \Gamma, A \wedge B, C \wedge D$ terminating with a $\wedge$-rule on $A \wedge B$ and, if ρ is any derivation of the same sequent terminating with a $\wedge$-rule on $C \wedge D$, they have the same set of leaves, namely:

$$\mathfrak{L}(\pi) = \sum_{i=1,\ldots,4} \mathfrak{L}(\pi_i) = \mathfrak{L}(\rho)$$

The other cases are similar.

A.3 Proof of Proposition 5.10

Suppose π is a derivation of $\vdash \Gamma$ in pLK_R, then by replacing every axiom rule by a ✠ rule we obtain a derivation π' of $\vdash \Gamma$ in $\mathrm{pLK}_R^{✠}$. Every $✠^{At}$ rule in π' is correct, since the sequent was introduced by an axiom rule in pLK_R.

Conversely, if π' is a derivation of $\vdash \Gamma$ in $\text{pLK}^{✠}_{R}$ such that $\mathfrak{L}(\pi')$ contains only correct sequent, then each sequent in $\mathfrak{L}(\pi')$ can be derived from an axiom rule in pLK_R. Therefore we obtain a derivation π of $\vdash \Gamma$ in pLK_R by replacing every $✠^{At}$ rule in π' by an axiom rule.

A.4 Proof of Lemma 5.12

Notice that, by definition of satisfiability of a sequent (and associativity of $\vee$), $\delta \vDash \Delta, A, B$ if and only if $\delta \vDash \Delta, A \vee B$.

In the case of the $\wedge$-rule, assume first that $\delta \vDash \Delta, A$ and $\delta \vDash \Delta, B$. Either there is satisfiable formula in Δ, either both A and B are satisfiable, therefore $\delta \vDash \Delta, A \wedge B$. Conversely, assume that $\delta \nvDash \Delta, A$, then all formulas in Δ are unsatisfiable and A is not satisfiable, therefore $A \wedge B$ is not satisfiable. We conclude that $\delta \nvDash \Delta, A \wedge B$. The lemma is then proved by simple induction.

A.5 Proof of Proposition 5.13

Suppose $\mathfrak{L}(\vdash \Gamma)$ contains only correct sequents, then for any valuation δ and sequents $\vdash \Delta$ in $\mathfrak{L}(\vdash \Gamma)$, $\delta \vDash \Delta$ from the definition of correct sequent. Then, by Lemma 5.12, $\delta \vDash \Gamma$.

Conversely, let us assume that $\mathfrak{L}(\vdash \Gamma)$ contains at least an incorrect sequent $\vdash P_1, \ldots, P_n, \neg Q_1, \ldots \neg Q_p$, such that for all integer $i \in [1, n]$ and $j \in [1, p]$, $P_i \not\equiv Q_j$. We can now take a valuation δ satisfying $\delta(P_i) = 0$ and $\delta(Q_j) = 1$. Then $\delta \nvDash P_1, \ldots, P_n, \neg Q_1, \ldots \neg Q_p$ and by Lemma 5.12 this means that $\delta \nvDash \Gamma$.

B Properties of System $\text{LK}^{✠}_{R}$

B.1 Proof of Lemma 5.19

Let π be such a derivation of the sequent $\vdash \Gamma$. Let $\mathfrak{L}(\pi)$ be the set of sequents introduced by ✠ rules in $\vdash \Gamma$, $\mathfrak{L}_c(\pi)$ the subset of $\mathfrak{L}(\pi)$ containing the sequents introduced by correct ✠ rules, and $\mathfrak{L}_a(\pi) = \mathfrak{L}(\pi) - \mathfrak{L}_c(\pi)$. By assumption, the sequents in $\mathfrak{L}_a(\pi)$ are introduced by admissible ✠ rules that are not correct. Hence there exists correct derivations π_i of $\vdash \Gamma_i$. Then, replacing the ✠ rules introducing the sequents $\vdash \Gamma_i$ in π by the derivations π_i, we obtain a derivation π' of $\vdash \Gamma$ extending π and containing only correct ✠ rules.

B.2 Proof of Theorem 5.21

Suppose we have a derivation of π in LK_R of a sequent $\vdash \Gamma$. Then, replacing every axiom rule by a ✠ yields a derivation π' of $\vdash \Gamma$ in $\mathrm{LK}_R^{✠}$. Moreover, the ✠ rules are all correct (since they were axiom rules in LK_R), hence admissible.

Conversely, suppose we have a derivation π' of a sequent $\vdash \Gamma$ in $\mathrm{LK}_R^{✠}$ that contains only admissible rules. Then, by Lemma 5.19 we can find a derivation π'' extending π' such that π'' contains only correct ✠ rules. Then, we can replace these ✠ rules by axiom rules to get a derivation π of $\vdash \Gamma$ in LK_R.

B.3 Proof of Lemma 5.23

Suppose π contains at least one ✠ rule introducing a sequent $\vdash \Gamma$ containing a formula B that is not an atom, a negation of an atom or an existential formula. Then the principal connective in B is either a $\wedge$, a $\vee$ or a $\forall$. Replacing the ✠ rule introducing $\vdash \Gamma$ by the rule introducing the principal connective and closing the derivation we obtain by ✠ rules then gives us a new derivation π_1 of π. After a finite number of iterations of this process, we obtain the wanted extension.

B.4 Proof of Lemma 5.25

Suppose now that the sequent $\vdash \Gamma$, $\Gamma = A_1, \ldots, A_m$, introduced by the non-admissible ✠ rule contains at least one quantifier. We fix an enumeration of the terms $t_1, \ldots, t_n, \ldots$ of the language and we will define an iterative process indexed by pairs (s, k_s) where s is a finite sequence of integers (the first step will be indexed by the null sequence of length m which will be written $(0)_m$) of length p_s and k_s is an integer in $[1, \ldots, p_s]$. The process we describe consists in extending the derivation by applying $\exists$ rules in a way that insures us that for all existential formula $\exists x A(x)$ and term t_i, there exists a step where the $\exists$ rule is used on the formula $A[t_i/x]$. To insure all terms appear at some point in the process we will use the enumeration but we need to keep track of the last term used for each existential formula. Moreover, applying a $\exists$ rule on a formula containing two existential connectives will produce new existential formulas on which we must apply the same procedure. The sequence will therefore keep track, for each existential formula, of the last term we used. Its length may vary, but due to our choice of existential rule it can only expand. The integer, on the other hand, will keep track of the last existential formula we decomposed, so that we can ensure that all formulas are taken into account.

First, let us write $A_1, \ldots, A_{p_{(0)_m}}$ the formulas in Γ that contain quantifiers. By Lemma 5.23 we can suppose, without loss of generality, that the ✠ rules in π are

simple. We will denote π by $\pi_{(0)_m}$, i.e. π will be the initial step of the process. The integer $k_{(0)_m}$ is defined to be 1, so we consider the derivation π_{A_1,t_1} obtained from π by replacing the ✠ rule introducing $\Gamma = \Delta, A_1, \ldots, A_p$ by the derivation consisting of a ✠ rule introducing $\Delta, A'_1[t_1/x], A_2, \ldots, A_{p_{(0)_m}}$ followed by an existential rule introducing A_1. It follows from Lemma 5.23 that this derivation can be extended to a simple derivation $\bar{\pi}_{A_1,t_1}$. Then, by the non-admissibility of the ✠ rule, this derivation contains at least one non-admissible ✠ rule introducing a sequent $\vdash \Gamma'$. Amongst the formulas of Γ are the all the formulas A_i for $1 \leqslant i \leqslant p$, but Γ' may contain more existential formulas. We thus denote by $A_1, \ldots, A_p$ the existential formulas of Γ'. We write $(0)^+_m = (1, 0, \ldots, 0)$ the sequence of length p: we thus obtained an extension $\pi_{(0)^+_m} = \bar{\pi}_{A_1,t_1}$ containing a non-admissible ✠ rule introducing a sequent $\Gamma_{(0)^+_m} = \bar{\Gamma}$. Defining $k_{(0)^+_m} = 2$, we arrived at the next step, indexed by $((0)^+_m, k_{(0)^+_m})$ and we can then iterate the process.

More generally, suppose we are at step (s, k_s) with $s = (s(0), \ldots, s(p))$: we have a simple derivation π_s with a non-admissible ✠ rule introducing a sequent $\Gamma_s = \Delta_s, A_1, \ldots, A_{p_s}$ (Δ_s contains only atoms and negations of atoms). We obtain a derivation $\pi_{A_{k_s},t_{s(k_s)+1}}$ by replacing the ✠ rule introducing Γ_s with the derivation:

$$\dfrac{\dfrac{}{\vdash \Delta_s, A_1, \ldots, A_{p_s}, A'_{k_s}[t_{s(k_s)+1}]}\,{\scriptstyle ✠}}{\vdash \Gamma_s}\,{\scriptstyle \exists}$$

This derivation $\pi_{A_{k_s},t_{s(k_s)+1}}$ can then be extended by Lemma 5.23 to a simple derivation $\bar{\pi}_{A_{k_s},t_{s(k_s)+1}}$ which contains a non-admissible ✠ rule. The sequent Γ' introduced by this rule contains all the formulas $A_1, \ldots, A_{p_s}$ and may contain additional existential formulas $A_{p_s+1}, \ldots, A_n$. Let s^+ to be the sequence of length n defined by $(s(0), \ldots, s(k_s - 1), s(k_s) + 1, s(k_s + 1), \ldots, s(p_s), 0, \ldots, 0)$, and:

$$k_{s^+} = \begin{cases} k_s + 1 \text{ if } \mathrm{k_s} + 1 \leqslant \mathrm{n} \\ 1 \qquad \text{otherwise} \end{cases}$$

Let us write $n = p_{s^+}$. We thus obtained the next step in the process, indexed by (s^+, k^+_s): a simple derivation $\pi_{s^+} = \bar{\pi}_{A_{k_s},t_{s(k_s)+1}}$ with a non-admissible ✠ rule introducing a sequent $\Gamma_{s^+} = \Gamma' = \Delta_{s^+}, A_1, \ldots, A_{p_{s^+}}$.

We claim that for all pairs (i, j) of natural numbers (different from 0), there is a step s in the process such that $s(i) = j$. We will write $\mathrm{len}(s)$ the length of a sequence s. Notice the formulas $A_{p_s+1}, \ldots, A_{p_{s^+}}$ are instantiations of an existential formulas A_i for $1 \leqslant i \leqslant p_s$ and therefore the number of existential connectives in a B_j is strictly less than the number of existential connectives in the corresponding A_i. We will show that the value of k_s returns to 1 in a finite number of steps using this remark. We will denote the number of existential connectives in a formula A by $\natural(A)$. Suppose that we are at a given step s such that $k_s = 1$ and write $o_s = (\max_{k_s < i \leqslant p_s} \natural(A_i), p_s - k_s)$. This pair somehow measures the number of steps one has to make before k_s returns to 1. Then, after $p_s - k_s$ steps in the process – let us write the resulting step as s^1, we have $k_{s^1} = p_s$ and $p_{s^1} - p_s$ new formulas, each one such that $\natural(A) < \max_{k_s < i \leqslant p_s} \natural(A_i)$. Therefore, the pair $o_{s^1} = (\max_{k_{s^1} < i \leqslant p_{s^1}} \natural(A_i), p_{s^1} - p_s)$.

Since $\max_{k_{s^1}<i\leqslant p_{s^1}} \natural(A_i) < \max_{k_s<i\leqslant p_s} \natural(A_i)$, we have $o_{s^1} < o_s$ in the lexicographical order and this is enough to show the claim.

B.5 Proof of Theorem 5.26

If the sequent $\vdash \Gamma$ introduced by the non-admissible ✠ rule does not contain any quantifiers, then the proof reduces to the proof of Proposition 5.13. Indeed, the derivation of $\vdash \Gamma$ in $pLK_R^{✠}$ is incorrect (if it were correct, it would contradict the assumption since any correct derivation in $pLK_R^{✠}$ is a correct derivation in $\mathrm{LK}_R^{✠}$), hence we can find a model $\mathscr{M}$ such that $\mathscr{M} \nVdash \Gamma$.

If $\vdash \Gamma$ contains existential formulas, we use Lemma 5.25 to obtain a sequence $(\pi_i)_{i\in\mathbf{N}}$ of extensions. From this sequence of extensions, one can obtain a sequence of sequents $\vdash \Gamma_i$ where for each i, there exists N such that $\vdash \Gamma_{i+1}$ is the premise of a rule whose conclusion is $\vdash \Gamma_i$ in all derivations π_j with $j \geqslant N$. Moreover, this sequence can be chosen so as to contain all instances of the subformulas of Γ. We now define a model whose base set is the set of terms. The interpretations of function symbols and constants are straightforward. The only thing left to define is the interpretation of predicates: if P is a n-ary predicate symbol, then $(t_1, \ldots, t_n)$ is in the interpretation of P if and only if $\forall i \geqslant 0$, $Pt_1 \ldots t_n \notin \Gamma_i$.

We can now check that $\mathscr{M} \nVdash \Gamma$. We chose A a formula in Γ and prove by induction on the size of the formula A that $\mathscr{M} \nVdash A$:

- if A is an atomic formula, then $\mathscr{M} \nVdash A$ by definition of the model;
- if $A = B \wedge C$, then there exists a sequent $\vdash \Gamma_i$ such that either $B \in \Gamma_i$ or $C \in \Gamma_i$. We suppose $B \in \Gamma_i$ without loss of generality. Then, by the induction hypothesis, we have $\mathscr{M} \nVdash B$, hence $\mathscr{M} \nVdash A$;
- if $A = B \vee C$, then there exists a sequent $\vdash \Gamma_i$ containing both B and C. By induction, these two formulas are not satisfied in the model $\mathscr{M}$, hence $\mathscr{M} \nVdash A$;
- if $A = \forall x B(x)$, then there is a sequent $\vdash \Gamma_i$ containing $B[y/x]$. By induction, $\mathscr{M} \nVdash B[y/x]$, hence $\mathscr{M} \nVdash A$;
- if $A = \exists x B(x)$, then for every term t there exists a sequent $\vdash \Gamma_i$ such that $B[t/x] \in \Gamma_i$. By the induction hypothesis, $\mathscr{M} \nVdash B[t/x]$. This being true for all term t, we conclude that $\mathscr{M} \nVdash A$.

This concludes the proof: since all formula $A \in \Gamma$ is such that $\mathscr{M} \nVdash A$, we have that $\mathscr{M} \nVdash \Gamma$.

References

Awodey, S., Reck, E.H.: Completeness and categoricity. Part I: nineteenth-century axiomatics to twentieth-century metalogic. Hist. Philos. Log. **23**(1), 1–30 (2002)

Bernays, P.: David Hilbert. In: Edwards, P. (ed.) Encyclopedia of Philosophy, vol. 3, pp. 496–505. MacMillan, New York (1967)

Bonnay, D.: Règles et signification: le point de vue de la logique classique. In: Joinet, J.-B. (ed.) Logique, Dynamique et Cognition, pp. 213–231. Publications de la Sorbonne, Paris (2007)
Bourbaki, N.: Theory of Sets. Hermann, Paris (1968)
Bourbaki, N.: Elements of the History of Mathematics. Springer, Berlin (1994)
Brouwer, L.E.J.: De Onbetrouwbaarheid der logische principes (The unreliability of the logical principles, English trans.). In: Heyting, A. (ed.) L.E.J. Brouwer Collected Works: Philosophy and Foundations of Mathematics, vol. 1, pp. 443–446. North-Holland, Amsterdam (1974/1908)
Di Cosmo, R., Miller, D.: Linear logic. In: Zalta, E.N. (ed.) Stanford Encyclopedia of Philosophy (2010). <http://plato.stanford.edu/archives/fall2010/entries/logic-linear>
Dowek, G.: From proof theory to theories theory. Manuscript (2010). https://who.rocq.inria.fr/Gilles.Dowek/Philo/leiden.pdf
Dowek, G., Hardin, T., Kirchner, C.: Theorem proving modulo. J. Autom. Reason. **31**(1), 33–72 (2003)
Dubucs, J.: Feasibility in logic. Synthese **132**(3), 213–237 (2002)
Dubucs, J., Marion, M.: Radical anti-realism and substructural logics. In: Rojszczak, A., Cachro, J., Kurczewski, G. (eds.) Philosophical Dimensions of Logic and Science. Selected Contributed Papers from the 11th International Congress of Logic, Methodology, and the Philosophy of Science, Krakòw, pp. 235–249. Kluwer, Dordrecht (2003)
Dummett, M.: Frege: Philosophy of Language. Duckworth, London (1973a)
Dummett, M.: The philosophical basis of intuitionistic logic. In: Dummett, M. (ed.) Truth and Other Enigmas, pp. 215–247. Duckworth, London (1973b/1978)
Dummett, M.: What is a theory of meaning? (II). In: Evans, G., McDowell, J. (eds.) Truth and Meaning: Essays in Semantics, pp. 67–137. Clarendon Press, Oxford (1976)
Dummett, M.: Elements of Intuitionism. Clarendon Press, Oxford (1977)
Dummett, M.: The Logical Basis of Metaphysics. Duckworth, London (1991)
Dummett, M.: The Seas of Language. Clarendon Press, Oxford (1993)
Genzten, G.: Untersuchungen über das logische Schliessen (Investigations into logical deduction, English trans.). In: Szabo, M.E. (ed.) The Collected Papers of Gerhard Gentzen, pp. 68–131. North-Holland, Amsterdam (1934–1935/1969)
Gentzen, G.: Neue Fassung des Widerspruchsfreiheitsbeweises für die reine Zahlentheorie (New version of the consistency proof for elementary number theory, English trans.). In: Szabo, M.E. (ed.) The Collected Papers of Gerhard Gentzen, pp. 252–308. North-Holland, Amsterdam (1938/1969)
Girard, J.-Y.: Proof Theory and Logical Complexity, vol. 1. Bibliopolis, Naples (1987)
Girard, J.-Y.: Linear logic. Theor. Comput. Sci. **50**(1), 1–101 (1987)
Girard, J.-Y.: Multiplicatives. In: Lolli, G. (ed.) Logic and Computer Science: New Trends and Applications, pp. 11–34. Rendiconti del seminario matematico dell'Università Politecnico di Torino, Torino (1988)
Girard, J.-Y.: Locus solum: from the rules of logic to the logic of rules. Math. Struct. Comput. Sci. **11**(3), 301–506 (2001)
Girard, J.-Y.: The Blind Spot. European Mathematical Society Publishing, Zürich (2011)
Girard, J.-Y.: Three lightings of logic. In: Ronchi Della Rocca, S. (ed.) Computer Science Logic 2013, pp. 1–23. Schloss Dagstuhl – Leibniz-Zentrum für Informatik/Dagstuhl Publishing, Wadern (2013)
Girard, J.-Y., Lafont, Y., Taylor, P.: Proofs and Types. Cambridge University Press, Cambridge (1989)
Hallett, M.: Hilbert and logic. In: Marion, M., Cohen, S. (eds.) Québec Studies in Philosophy of Science, vol. 1, pp. 135–187. Kluwer, Dordrecht (1995)
Hempel, C.G.: Geometry and empirical science. Am. Math. Mon. **52**(1), 7–17 (1945)
Heyting, A.: Axiomatic method and intuitionism. In: Bar-Hillel, Y. (ed.) Essays on the Foundations of Mathematics: Dedicated to A.A. Fraenkel on his Seventieth Anniversary, pp. 237–247. Magnes Press, Jerusalem (1962)
Hilbert, D.: Logische Principien des mathematischen Denkens, Ms. Vorlesung SS 1905, annotated by E. Hellinger, Bibliothek des Mathematischen Seminars, Universität Göttingen (1905)

Hindley, J.R., Seldin, J.P.: Lambda-Calculus and Combinators: An Introduction. Cambridge University Press, Cambridge (2008)
Hintikka, J.: The Game of Language: Studies in Game-Theoretical Semantics and its Applications, in Collaboration with J. Kulas. Kluwer, Dordrecht (1983)
Hintikka, J.: What is the axiomatic method? Synthese **183**(1), 69–85 (2011)
Hjortland, O.: Harmony and the context of deducibility. In: Dutilh Novaes, C., Hjortland, O. (eds.) Insolubles and Consequences: Essays in Honour of Stephen Read, pp. 105–117. College Publications, London (2012)
Hyland, J.M.E.: Proof theory in the abstract. Ann. Pure Appl. Log. **114**(1–3), 43–78 (2002)
Kamlah, W., Lorenzen, P.: Logische Propädeutik, 2nd edn. Metzler, Stuttgart/Weimar (1972)
Keiff, L.: Dialogical logic. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy (2009). http://plato.stanford.edu/archives/sum2011/entries/logic-dialogical/
Kreisel, G.: Mathematical significance of consistency proofs. J. Symb. Log. **23**(2), 155–182 (1958)
Kreisel, G.: Foundations of intuitionistic mathematics. In: Nagel, E., Suppes, P., Tarski, A. (eds.) Logic, Methodology and Philosophy of Science: Proceedings of the 1960 International Congress, pp. 198–210. Stanford University Press, Stanford (1962)
Kreisel, G.: Proof theory: some personal recollections. In: Takeuti, G. (ed.) Proof Theory, pp. 395–405. North Holland, Amsterdam (1987)
Krivine, J.-L.: Lambda Calculus, Types and Models. Ellis Horwood, Hemel Hempstead (1993)
Krivine, J.-L.: Dependent choice, 'quote' and the clock. Theor. Comput. Sci. **308**(1–3), 259–276 (2003)
Lecomte, A., Quatrini, M.: Figures of dialogue: a view from ludics. Synthese **183**(Supplement 1), 279–305 (2011a)
Lecomte, A., Quatrini, M.: Ludics and rhetorics. In: Lecomte, A., Tronçon, S. (eds.) Ludics, Dialogue and Interaction. PRELUDE Project 2006–2009: Revised Selected Papers. Lecture Notes in Artificial Intelligence, vol. 6505, pp. 32–59. Springer, Berlin (2011b)
Lorenzen, P., Lorenz, K.: Dialogische Logik. Wissenschaftliche Buchgesellschaft, Darmstadt (1978)
Lorenzen, P., Schwemmer, O.: Konstruktive Logik, Ethik und Wissenschaftstheorie, 2nd edn. Bibliographisches Institut, Mannheim (1973)
Marion, M.: Radical anti-realism, Wittgenstein and the length of proofs. Synthese **171**(3), 419–432 (2009)
Marion, M.: Game semantics and the manifestation thesis. In: Rahman, S., et al. (eds.) The Realism-Antirealism Debate in the Age of Alternative Logics, pp. 141–168. Springer, Berlin (2012)
Martin-Löf, P.: Truth of a proposition, evidence of a judgment, validity of a proof. Synthese **73**(3), 407–420 (1987)
Martin-Löf, P.: On the meanings of the logical constants and the justifications of the logical laws. Nord. J. Philos. Log. **1**(1), 11–60 (1996)
Naibo, A.: Le statut dynamique des axiomes. Des preuves aux modèles. Ph.D. thesis, Université Paris 1 Panthéon-Sorbonne (2013)
Naibo, A., Petrolo, M.: Are uniqueness and deducibility of identicals the same? Theoria **81**, 143–181 (2015)
Naibo, A., Petrolo, M., Seiller, T.: A computational analysis of logical constants. In: Workshop on Logical Constants, Ljubljana, 7–12 Aug 2011. http://lumiere.ens.fr/~dbonnay/files/conference/LC/NPS.pdf
Negri, S., von Plato, J.: Cut elimination in the presence of axioms. Bull. Symb. Log. **4**(4), 418–435 (1998)
Negri, S., von Plato, J.: Structural Proof Theory. Cambridge University Press, Cambridge (2001)
Negri, S., von Plato, J.: Proof Analysis: A Contribution to Hilbert's Last Problem. Cambridge University Press, Cambridge (2011)
Paoli, F.: Substructural Logics: A Primer. Kluwer, Dordrecht (2002)

Pasch, M.: Begriffsbildung und Beweis in der Mathematik [Concepts and proofs in mathematics]. In: Pollard, S. (English trans.) (ed.) Essays on the Foundations of Mathematics by Moritz Pasch, pp. 183–203. Springer, Berlin (2010/1925)

Pereira, L.C.: On the estimation of the length of normal derivations. Ph.D. thesis, Stockholm University (1982)

von Plato, J.: Translations from natural deduction to sequent calculus. Math. Log. Q. **49**(5), 435–443 (2003)

von Plato, J.: In the shadows of the Löwenheim-Skolem theorem: early combinatorial analyses of mathematical proofs. Bull. Symb. Log. **13**(2), 189–225 (2007)

Poggiolesi, F.: Gentzen Calculi for Modal Propositional Logic. Springer, Berlin (2011)

Prawitz, D.: Natural Deduction: A Proof-Theoretical Study. Almqvist & Wiksell, Stockholm (1965)

Prawitz, D.: Towards a foundation of a general proof theory. In: Suppes, P., et al. (eds.) Logic, Methodology and Philosophy of Science IV. Proceedings of the Fourth International Congress for Logic, Methodology and Philosophy of Science, Bucharest, 1971, pp. 225–250. North-Holland, Amsterdam (1973)

Proclus, Morrow, G.R. (ed.) A Commentary on the First Book of Euclid's Elements. Princeton University Press, Princeton (1970)

Rahman, S., Clerbout, N.: Constructive type theory and the dialogical approach to meaning. In: Marion, M., Pietarinen, A.-V. (eds.) Games, Game Theory and Game Semantics. The Baltic International Yearbook of Cognition, Logic and Communication, vol. 8, pp. 1–72 (2013). doi:10.4148/1944-3676.1077

Rahman, S., Keiff, L.: On how to be a dialogician. In: Vanderverken, D. (ed.) Logic, Thought and Action, pp. 359–408. Kluwer, Dordrecht (2004)

Redmond, J., Fontaine, M.: How to Play Dialogues: An Introduction to Dialogical Logic. College Publications, London (2011)

Rückert, H.: Dialogues as a Dynamic Framework for Logic. College Publications, London (2011)

Schütte, K.: Ein System des verknüpfenden Schliessens. Archiv für mathematische Logic und Grundlagenforschung **2**(2–4), 55–67 (1956)

Seiller, T.: Interaction graphs: additives. Ann. Pure Appl. Log. **167**, 95–154 (2016)

Seiller, T.: Interaction graphs: multiplicatives. Ann. Pure Appl. Log. **163**(12), 1808–1837 (2012b)

Seiller, T.: Logique dans le facteur hyperfini: Géométrie de l'interaction et complexité. Ph.D. thesis, Université Aix-Marseille (2012c)

Seiller, T.: Interaction graphs: exponentials (2013, Submitted). arXiv:1312.1094

Seiller, T.: Interaction graphs: graphings (2014, Submitted). arXiv:1405.6331

Sørensen, M.H., Urzyczyn, P.: Lectures on the Curry-Howard Isomorphism. Elsevier, Amsterdam (2006)

Sundholm, G.: Constructions, proofs and the meaning of logical constants. J. Philos. Log. **12**(2), 151–172 (1983)

Sundholm, G.: Questions of proof. Manuscrito **16**(2), 47–70 (1993)

Sundholm, G.: Implicit epistemic aspects of constructive logic. J. Log. Lang. Inf. **6**(2), 191–212 (1997)

Sundholm, G.: Proofs as acts and proofs as objects: some questions for Dag Prawitz. Theoria **64**(2–3), 187–216 (1998)

Sundholm, G.: Antirealism and the roles of truth. In: Niiniluoto, I., Simonen, M., Woleński, J. (eds.) Handbook of Epistemology, pp. 437–466. Kluwer, Dordrecht (2004)

Tennant, N.: Inferentialism, logicism, harmony, and a counterpoint. In: Miller, A. (ed.) Essays for Crispin Wright: Logic, Language and Mathematics, vol. 2. Oxford University Press, Oxford (2012, to appear)

Troelstra, A.S., Schwichtenberg, H.: Basic Proof Theory, 2nd edn. Cambridge University Press, Cambridge (2000)

Troelstra, A.S., van Dalen, D.: Constructivism in Mathematics, vol. 2. North-Holland, Amsterdam (1988)
Usberti, G.: Risposta a Casalegno. Lingua e Stile **32**(3), 529–536 (1997)
Wang, H.: Popular Lectures on Mathematical Logic. van Nostrand, New York (1981)
Wansing, H.: The idea of a proof-theoretic semantics and the meaning of the logical operations. Studia. Logica. **64**, 3–20 (2000)
Wittgenstein, L.: Philosophical Investigations. G.E.M. Anscombe and R. Rhees (eds.), G.E.M. Anscombe (trans.). Blackwell, Oxford (1953)
Wittgenstein, L.: Bemerkungen über die Grundlagen der Mathematik. In: von Wright, G.H., Rhess, R., Anscombe, G.E.M. (eds.) Remarks on the Foundations of Mathematics (English trans. by Anscombe, G.E.M.). Basil Blackwell, Oxford (1956)

Part II
The Dynamics of Knowledge II: Epistemology, Games, and Dynamic Epistemic Logic

Chapter 6
A Dynamic Analysis of Interactive Rationality

Eric Pacuit and Olivier Roy

Abstract Epistemic game theory has shown the importance of informational contexts to understand strategic interaction. We propose a general framework to analyze how such contexts may arise. The idea is to view informational contexts as the fixed points of iterated, rational responses to incoming information about the agents' possible choices. We discuss conditions under which such fixed points may exist. In the process, we generalize existing rules for information updates used in the dynamic epistemic logic literature. We then apply this framework to weak dominance. Our analysis provides a new perspective on a well known problem with the epistemic characterization of iterated removal of weakly dominated strategies.

Keywords Game theory • Dynamic epistemic logic • Rationality • Update • Fixed points • Admissibility

6.1 Introduction

A crucial assumption underlying classical game-theoretic analyses is that there is *common knowledge* that all the players are *rational*. Rationality, here, is understood in the decision-theoretic sense: The players' choices are *optimal* according to some choice rule (such as maximizing subjective expected utility). Recent work in *epistemic game theory* has focused on developing sophisticated mathematical models to study the implications of assuming that all the players are rational and that this is commonly known (or commonly believed).[1] However, if common

[1] See Perea (2012), Dekel and Siniscalchi (2015), and Pacuit and Roy (2015) for surveys of this literature.

E. Pacuit (✉)
Department of Philosophy, University of Maryland, College Park, MD, USA
e-mail: epacuit@umd.edu

O. Roy
Institut für Philosophie, Universität Bayreuth, Bayreuth, Germany
e-mail: olivier.roy@uni-bayreuth.de

J. Redmond et al. (eds.), *Epistemology, Knowledge and the Impact of Interaction*,
Logic, Epistemology, and the Unity of Science 38, DOI 10.1007/978-3-319-26506-3_6

knowledge of rationality is to have an "explanatory" role in the analysis of a game-theoretic situation, then it is not enough to simply *assume* that it has obtained in an informational context of a game. It is also important to describe how the players were able to arrive at this crucial state of information.[2]

There is a growing body of literature focused on analyzing games in terms of the "process of deliberation" that leads the players to select their component of a rational outcome. Many different frameworks have been used to represent this process of deliberation:

1. John Harsanyi's "tracing procedure" identifies a unique Nash equilibrium in any finite strategic game that is the limit of a sequence of Nash equilibria from a related set of strategic games. Harsanyi thought of the tracing procedure as "a mathematical formalization of the process by which rational players coordinate their choices of strategies" (Harsanyi 1975).
2. Brian Skyrms assumes that players deliberate by calculating their subjective expected utility and then use the results of their calculations to adjust their probabilities about what they are going to do and what they expect their opponents to do (Skyrms 1990).
3. Robin Cubitt and Robert Sugden apply David Lewis's "common modes of reasoning" to game-theoretic situations. They describe the players' process of deliberation as an iterative procedure for classifying strategies (Cubitt and Sugden 2011, 2014).
4. Johan van Benthem and colleagues use ideas from dynamic epistemic logic to characterize solution concepts as fixed points of iterated "(virtual) rationality announcements" (Baltag et al. 2009; van Benthem 2014).

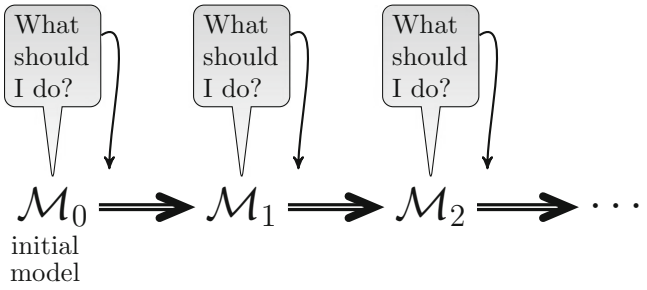
Although the details of these frameworks[3] are different, they share a common line of thought: The rational outcomes of a game are arrived at through a process in which each player settles on an optimal choice given her evolving beliefs about her own and her opponents' choices. This is not intended to be a formal account of the players' *practical reasoning* in game situations. Rather, the goal is to describe deliberation in terms of a sequence of belief changes about what the players are doing and what their opponents may be thinking. The general conclusion is that the rational outcomes of a game depend not only on the structure of the game and the players' initial beliefs, but also on which dynamical rule the players are using to update their inclinations and beliefs, and what exactly is commonly known about the process of deliberation. For instance, the outcomes of Harsanyi's tracing procedure and Skryms's model of dynamic deliberation are qualitatively similar: Both procedures lead players to choose their component of a Nash equilibrium. However, in Skyrms's model, the rate of convergence depends on the players'

[2]David Lewis already appreciated this general point about common knowledge when he first formulated his notion of common knowledge (Lewis 1969). See Cubitt and Sugden (2003) for an illuminating discussion and a reconstruction of Lewis' notion of common knowledge, with applications to game theory.

[3]See Pacuit (2015) for an extensive discussion of these different models of deliberation in games.

initial beliefs and the dynamical rule changing the players' inclinations during deliberation; and this, in turn, suggests a more refined analysis of Nash equilibrium (Skyrms 1990, pp. 154–158).

There are two key components of the above models of deliberation in games. The first component is a formal representation of the players' *state of indecision*. This is intended to be a "snapshot" of the players' *inclinations* about what they are going to choose and their *beliefs* about their opponents' choices and beliefs during the process of deliberation. The second component is the dynamical rule that governs the changes in the players' state of indecision. The general idea is that, at each stage of the deliberation, the players determine which of their available strategies are "optimal" and which they ought to avoid. Typically, it is assumed that the players are guided by some decision-theoretic choice rule, such as maximizing expected utility or avoiding dominated strategies. Using the information about the players' own choices and what they expect their opponents to do, the players' state of indecision is *transformed* according to some fixed dynamical rule. The picture to keep in mind is:



Deliberation concludes when the players reach a fixed point in the above process. The central question is: What types of transformations match different game-theoretic analyses?

In this paper, we develop a model of deliberation and characterize whether players will reason to specific informational contexts (Sect. 6.2). We then apply this framework to issues surrounding the epistemic characterization of *iterated elimination of weakly dominated strategies* (IEWDS), aka *iterated admissibility* (Sect. 6.3). Our approach builds on earlier work that describes deliberation in games in terms of (virtual) rationality announcements (van Benthem 2007; Baltag et al. 2009; Baltag and Smets 2009; van Benthem and Gheerbrant 2010).

6.2 Belief Dynamics for Strategic Games

The main idea of this paper is to understand well-known solution concepts not in terms of fixed informational contexts—for instance, models (e.g., type spaces or epistemic models) satisfying rationality and common belief of rationality—but,

rather, as a result of a dynamic, interactive deliberation process. It is important to note that the goal is *not* to represent some type of "pre-play communication" or some form of "cheap talk". Instead, the goal is to represent the process of *rational deliberation* that takes the players from the *ex ante* stage to the *ex interim* stage of decision making. In this section, we introduce our framework, incorporating ideas from the extensive literature on dynamic logics of belief revision (van Benthem 2010; Baltag and Smets 2009) and recent work on the reasoning-based approach to game theory found in Cubitt and Sugden (2011, 2014).

6.2.1 Strategic Games and Game Models

A finite strategic game is a tuple $G = \langle N, \{S_i\}_{i\in N}, \{u_i\}_{i\in N}\rangle$, where N is a finite set of players; for each $i \in N$, S_i is a finite set of actions (also called strategies) for player i; and for each $i \in N$, $u_i : \Pi_{i\in N} S_i \to \mathbb{R}$ is a utility function assigning real numbers to each outcome of the game.[4] A **strategy profile** is a tuple $\vec{s} \in \Pi_{i\in N} S_i$, specifying an action for each player. Following standard game-theoretic notation, we write $\vec{s}_{-i} \in \Pi_{j\in N-\{i\}} S_i$ for a sequence of actions for all players except i. For simplicity, we assume that the outcomes of the game G are identified with the set of strategy profiles $S = \Pi_{i\in N} S_i$.

A **game model** describes the players' *hard* and *soft* information about the possible outcomes of the game. The models that we use in this paper are standard in the belief revision literature: a non-empty set of states, where each state is associated with a possible outcome of the game, and a single relation $\preceq$ on W representing the players' (common) initial plausibility ordering. Originally used as a semantics for conditionals (cf. Lewis 1973), these *plausibility models* have been extensively used by logicians (van Benthem 2004, 2010; Baltag and Smets 2009), game theorists (Board 2004) and computer scientists (Boutilier 1992; Lamarre and Shoham 1994) to represent rational agents' (all-out) beliefs. Thus, we take for granted that they provide natural models of (multiagent) beliefs and focus on how they can be used to represent "rational deliberation" in a game situation. The formal definition of a game model is as follows.

Definition 6.2.1 (Strategy Functions). Suppose that W is a non-empty set of states, and $G = \langle N, \{S_i\}_{i\in N}, \{u_i\}_{i\in N}\rangle$ is a finite strategic game. A **strategy function** on W for G is a function $\sigma : W \longrightarrow S$ assigning strategy profiles to each state. To simplify notation, we write $\sigma_i(w)$ for $(\sigma(w))_i$ (similarly, write $\sigma_{-i}(w)$ for the sequence of strategies of all players except i).

[4] We assume that the reader is familiar with the basic concepts of game theory (e.g., strategic games and various solution concepts such as iterated removal of strictly/weakly dominated strategies). Consult Leyton-Brown and Shoham (2008) for an introduction to game theory.

Definition 6.2.2 (Game Model). Suppose that $G = \langle N, \{S_i\}_{i \in N}, \{u_i\}_{i \in N} \rangle$ is a finite strategic game. A **model** of G is a tuple $\mathcal{M}_G = \langle W, \preceq, \sigma \rangle$, where W is a non-empty set; $\preceq$ is a connected, reflexive, transitive and well-founded[5] relation on W; and σ is a strategy function on W for G. Subsets of W are called **events**, or **propositions**.

Note that there is only one plausibility ordering in the above model; yet we are interested in games with more than one player. There are different ways to interpret the fact that there is only one plausibility ordering. First, we can take the perspective of a single player thinking about what she is going to choose in the game. Alternatively, we can think of the model as describing a stage of the rational deliberation of *all* the players, starting from a situation in which the players have the same beliefs (i.e., there is a *common prior*). The players' private beliefs, *given their actual choice of strategy*, can be defined using conditional beliefs.[6] We first need some notation. For $\emptyset \neq X \subseteq W$, let $Min_{\preceq}(X) = \{v \in X \mid v \preceq w \text{ for all } w \in X\}$ be the set of minimal elements of X according to $\preceq$. This set contains the *most plausible* states in X.

Definition 6.2.3 (Belief and Conditional Belief). Suppose that $\mathcal{M}_G = \langle W, \preceq, \sigma \rangle$ is a model of a finite strategic game G. For all subsets E and F of W, E is believed conditional on F is defined as follows:

$$B(E \mid F) = \{w \mid Min_{\preceq}(F) \subseteq E\}.$$

We also write $B^F(E)$ for $B(E \mid F)$. If $w \in B^F(E)$, then we say that "E is believed conditional on F at w". Also, we say that E is **believed** in $\mathcal{M}_G$ if E is believed conditional on W. Thus, E is believed when $Min_{\preceq}(W) \subseteq E$.

Of course, the game models from Definition 6.2.2 can be (and have been: see Baltag and Smets 2009; van Benthem 2010) be extended to include plausibility orderings for each player, state-dependent plausibility ordering(s), explicit relations representing the players' knowledge about the game situation, and other notions of beliefs (e.g., *strong belief* or *robust belief*). To keep things simple, we focus on models with a single plausibility ordering.

[5] Well-foundedness is only needed to ensure that for any set X, the set of minimal elements in X is nonempty. This is important only when W is infinite – and there are ways around this in current logics. Moreover, the condition of connectedness can also be lifted, but we use it here for convenience.

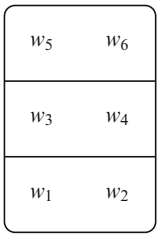
[6] A similar idea is found in standard models of differential information from the economics literature. In such models, it is assumed that there is a prior probability measure describing the players' initial beliefs (often it is the same probability measure for all the players). The players' *posterior probabilities* are defined by conditioning their prior probability measure on their private information (typically represented by some partition over the set of states).

6.2.2 A Primer on Belief Dynamics

We are not interested in game models per se, but, rather, how a game model changes during the process of rational deliberation. The type of changes we are interested in is how a model $\mathcal{M}_G$ of a game G incorporates new information about what the players *should* do (according to a some decision-theoretic choice rule). As is well known from the belief revision literature, there are many ways to transform a plausibility model given some new information (Rott 2006). We do not have the space to survey this entire literature here (see van Benthem (2010) and Pacuit (2013) for modern introductions). Instead, we sketch some key ideas.

The general approach is to define a way of *transforming* a game model $\mathcal{M}_G$ given an event E. That is, we will define functions τ that map game models and events to game models. For each game model $\mathcal{M}_G$ and event E, we write $\mathcal{M}_G^{\tau(E)}$ for $\tau(\mathcal{M}_G, E)$. So, given a model $\mathcal{M}_G$ of a game G and an event E describing what the players (might/should/will) do, $\mathcal{M}_G^{\tau(E)}$ is the updated game model, taking this information into account. Different definitions of τ represent the different attitudes that an agent can have towards the incoming information.

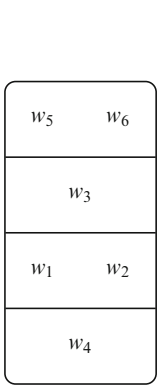
We start with an illustrative example. Suppose that $\mathcal{M}_G = \langle W, \preceq, \sigma\rangle$ is a game model in which $W = \{w_1, w_2, w_3, w_4, w_5, w_6\}$, and $\preceq$ is defined as follows: $w_1 \sim w_2 \prec w_3 \sim w_4 \prec w_5 \sim w_6$, where $w \prec v$ means $w \preceq v$ and $v \npreceq w$ and $w \sim v$ means $w \preceq v$ and $v \preceq w$. This game model is pictured as follows:



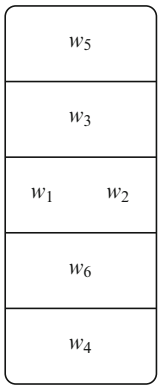
The first transformation that we discuss is the well-known *public announcement* operation (Plaza 1989; Gerbrandy 1999), denoted by !. This operation assumes that the players considers the source of the new information E *infallible*, ruling out any states not contained in E. That is, the updated model $\mathcal{M}_G^{!(E)}$ is $\langle E, \preceq', \sigma'\rangle$, where $\preceq' = \preceq \cap E$ and σ' is σ restricted to E.

Two other transformations have been widely discussed in the belief revision literature. For these transformations, the players do *trust* the source of the new information, though they do not treat the source as infallible. Perhaps the most ubiquitous transformation is *conservative upgrade* ($\uparrow(E)$), which lets the players only tentatively accept the incoming information E by making the most plausible E-states the new minimal set and keeping the old plausibility ordering the same on all other states. A second transformation is *radical upgrade* ($\Uparrow(E)$), which moves

all the states in E below all the other states and, otherwise, keeps the plausibility ordering the same. The results of these two operations with $E = \{w_4, w_6\}$ on the above model $\mathcal{M}_G$ are:



$\mathcal{M}_G^{\uparrow(E)}$



$\mathcal{M}_G^{\Uparrow(E)}$

These transformations satisfy a number of interesting logical principles (van Benthem 2010) that we do not discuss in this paper.

We are interested in using these transformations to model the players' process of deliberation in a game. Given a game model (viewed as describing one stage of the deliberation process), the players determine which options are "rationally permissible" and which options the players ought to avoid (as specified by some decision-theoretic choice rule). Given this information, the players use one of the above transformations to change the game model. In this new game model, the players reconsider what they should do leading to another transformation. The main question is: does this process *stabilize*?

The answer to this question will depend on a number of factors. The general picture is

$$\mathcal{M}_0 \overset{\tau(D_0)}{\Longrightarrow} \mathcal{M}_1 \overset{\tau(D_1)}{\Longrightarrow} \mathcal{M}_2 \overset{\tau(D_2)}{\Longrightarrow} \cdots \overset{\tau(D_n)}{\Longrightarrow} \mathcal{M}_{n+1} \Longrightarrow \cdots,$$

where each D_i is some event and τ is a model transformer (e.g., public announcement, radical upgrade or conservative upgrade). Two questions are important for the analysis of this process. First, what type of transformations are the players using? For example, if τ is the public announcement transformation, then it is not hard to see that, for purely logical reasons, this process must eventually stop at a limit model (see Baltag and Smets (2009) for a discussion and proof). Second, where do the propositions D_i come from? To see why this matters, consider the situation in which you iteratively perform a radical upgrade with E and $\overline{E}$ (the complement of

E). Of course, this sequence of upgrades never stabilizes. However, in the context of reasoning about what to do in a game situation, this situation may not arise because of special properties of the choice rule that is being used to generate the events D_i.

6.2.3 Categorizing Strategies

Any sequence of game models can be viewed as the stages of a process of deliberation in the underlying game. We are interested primarily in sequences of game models that are generated by some fixed belief transformation (such as a public announcement, a conservative upgrade, or a radical upgrade). However, it is not enough to simply fix an initial game model and some model transformation to represent the players' deliberation about what they are going to do in a game. We also need a way to define the events used to update the models at each stage of the deliberation. These events should specify, for each player, which actions are "rationally permissible" and which actions they should avoid. In this section, we discuss the key features of any general method that can be used to identify the events that will serve as input to the model transformation at each stage of the deliberation.

We start with two general observations about decision making in games to motivate the definitions in this section. The first observation is that, in general, there are no rational principles of "rational" decision making (under ignorance or uncertainty) that *always* recommend a *unique* choice.[7] In particular, it is not hard to find a game and a game model where there is at least one player without a *unique* "rational choice". Making use of a well-known distinction of Ullmann-Margalit and Morgenbesser (1977), the assumption that all players are rational can help determine which options the player ought to *choose*. However, since nothing distinguishes between these on rationality grounds alone, the player is left to *pick* any of the rationally permissible options.[8]

The second observation is that we do not intend our model of deliberation to directly represent the *practical reasoning* leading to the players' decision about what to do in a game situation. In fact, we do not directly represent any formal model of practical reasoning. Instead, we treat practical reasoning as a "black box" and focus on general *choice rules* that are intended to describe the outcome of the players' practical reasoning. More generally, following Cubitt and Sugden (2014), we assume that during each stage of deliberation, the players can *categorize* their available actions. To make this precise, we need some notation:

[7] Consult any textbook on decision theory, such as Peterson (2009), for evidence of this fact.

[8] See Roy et al. (2014) and Anglberger et al. (2015) for a discussion on the rational obligations and permissions in games.

Definition 6.2.4 (Strategies in Play). Suppose that $G = \langle N, \{S_i\}_{i\in N}, \{u_i\}_{i\in N}\rangle$ is a finite strategic game and $\mathcal{M}_G = \langle W, \preceq, \sigma\rangle$ is a model of G. For each $i \in N$, the strategies **in play for** i is the set

$$S_{-i}(\mathcal{M}_G) = \{s_{-i} \in \Pi_{j\neq i} S_j \mid \text{ there is a } w \in Min_{\preceq}(W) \text{ such that } \sigma_{-i}(w) = s_{-i}\}.$$

The set $S_{-i}(\mathcal{M}_G)$ contains the strategies that player i still *believes* are possible at some stage of the deliberation process represented by the model $\mathcal{M}_G$. Given these beliefs, we assume that each player can *categorize* her available options:

Definition 6.2.5 (Categorization). Let $G = \langle N, \{S_i\}_{i\in N}, \{u_i\}_{i\in N}\rangle$ be a strategic game and $\mathcal{M}_G = \langle W, \preceq, \sigma\rangle$ a model of G. A **categorization** for player i in $\mathcal{M}_G$ is a pair $\mathbf{S}_i(\mathcal{M}_G) = (S_i^+, S_i^-)$ where $S_i^+ \cup S_i^- \subseteq S_i$, $S_i^+ \cap S_i^- = \emptyset$, and

$$(*) \qquad \text{for each } a \in S_i, \text{if there is no } v \in W \text{ with } \sigma_i(v) = a, \text{then } a \in S_i^-.$$

If $\mathbf{S}_i(\mathcal{M}_G) = (S_i^+, S_i^-)$, we write $\mathbf{S}_i^+(\mathcal{M}_G)$ for S_i^+ and $\mathbf{S}_i^-(\mathcal{M}_G)$ for S_i^-. Also, we write $\mathbf{S}(\mathcal{M}_G)$ for the sequence of categorizations $(\mathbf{S}_i(\mathcal{M}_G))_{i\in N}$.

The intended interpretation is that player i ought to pick from among the strategies in $\mathbf{S}_i^+(\mathcal{M}_G)$ and ought to avoid any strategy in $\mathbf{S}_i^-(\mathcal{M}_G)$. The strategies in $S_i - (\mathbf{S}_i^+(\mathcal{M}_G) - \mathbf{S}_i^-(\mathcal{M}_G))$ have not yet been categorized. These are the strategies that player i needs to think more about before categorizing. Condition $(*)$ in the above definition ensures that players will not choose any strategy that has been completely ruled out. Note that, in general, a categorization need not be a partition of player i's strategies (i.e., $\mathbf{S}_i^+(\mathcal{M}_G) \cup \mathbf{S}_i^-(\mathcal{M}_G) \neq S - i$). See Cubitt and Sugden (2011) for an example of such a categorization. However, many of the familiar choice rules found in the game theory literature lead to categorizations that do form a partition. Two standard examples are weak and strong dominance: Let $G = \langle N, \{S_i\}_{i\in N}, \{u_i\}_{i\in N}\rangle$ be a strategic game and $\mathcal{M}_G$ a model of G. Then:

Strong Dominance (pure strategies): For each $i \in N$, $\mathbf{SD}_i(\mathcal{M}_G) = (S_i^+, S_i^-)$ is defined as follows: For all $a \in S_i$,

$$a \in S_i^- \text{ iff there is } b \in S_i \text{ such that for all } s_{-i} \in S_{-i}(\mathcal{M}_G),\ u_i(s_{-i}, b) > u_i(s_{-i}, a),$$

and $S_i^+ = S_i - S_i^-$.

Weak Dominance (pure strategies): For each $i \in N$, $\mathbf{WD}_i(\mathcal{M}_G) = (S_i^+, S_i^-)$ is defined as follows: For all $a \in S_i$,

$$a \in S_i^- \text{iff there is } b \in S_i \text{such that for all } s_{-i} \in S_{-i}(\mathcal{M}_G),\ u_i(s_{-i}, b) \geq u_i(s_{-i}, a)$$

and there is some

$$s_{-i} \in S_{-i}(\mathcal{M}_G) \text{ such that } u_i(s_{-i}, b) > u_i(s_{-i}, a),$$

and $S_i^+ = S_i - S_i^-$.

Both of the above definitions can be modified to cover strict/weak dominance by *mixed strategies*, but we leave issues about how to incorporate probabilities into the framework sketched in this paper for another time.

We conclude this section by defining when a game model *incorporates* a categorization. Suppose that $\mathcal{M}_G = \langle W, \preceq, \sigma \rangle$ is a game model for G, and $\mathbf{S}(\mathcal{M}'_G)$ is a categorization for a game model $\mathcal{M}'_G$. We say that $\mathcal{M}_G$ **incorporates** $\mathbf{S}(\mathcal{M}'_G)$ provided that for all $i \in N$:

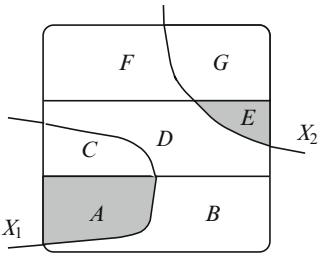
- If $a \in \mathbf{S}_i^+(\mathcal{M}'_G)$, then there is some $w \in Min_{\preceq}(W)$ such that $\sigma_i(w) = a$.
- If $a \in \mathbf{S}_i^-(\mathcal{M}'_G)$, then there is no $w \in Min_{\preceq}(W)$ such that $\sigma_i(w) = a$.

Thus, a model $\mathcal{M}_G$ incorporating a categorization $(S_i^+, S_i^-)_{i \in N}$ implies that (1) for each $a \in S_i^+$, the players do not believe that i will not play a; and (2) for each $a \in S_i^-$, players believe that i will not play a.

6.2.4 Generalized Belief Transformations

An important feature of a categorization is that more than one strategy may be "rationally permissible" for a player. This means that the information the players gain from a categorization should be represented by a *set* of events rather than a single event. Each event in this set describes the outcomes of the game that result from assuming that each player picks a rationally permissible strategy. In this section, we show how to generalize the model transformations introduced in Sect. 6.2.2 to accept finite sets of events as inputs.

Suppose that $\{E_1, \ldots, E_k\}$ is a set of events for game model $\mathcal{M}_G$. The generalization of the public announcement transformation is straightforward: $!\{E_1, \ldots, E_k\}) = !(E_1 \cup E_2 \cup \cdots E_k)$. The generalizations of the conservative and radical upgrade is more subtle. To see the difficulty, consider the game model pictured below with two events, X_1 and X_2:



The sets A, B, C, D, E, F and G denote all the different subsets of states (so, $W = A \cup B \cup C \cup D \cup E \cup F \cup G$). The plausibility ordering runs from the top to the bottom. So, for instance, the states in $A \cup B$ are the most plausible overall, and all states within $A \cup B$ are equiplausible. A conservative upgrade with $X_1 \cup X_2$ results in the following modification of the above plausibility ordering:

$$A \prec B \prec C \cup D \cup E \prec F \cup G,$$

where for sets X and Y, we write $X \prec Y$ when for all $w \in X, v \in Y, w \prec v$. However, suppose that X_1 and X_2 describe two different outcomes of the game. Furthermore, in each of the outcomes, assume that the players pick a rationally permissible action. The result of the radical upgrade with $X_1 \cup X_2$ is that the players come to believe that the outcome of the game will be as described by X_1. The beliefs are the same after a radical upgrade with $X_1 \cup X_2$, though the resulting plausibility ordering is different.

However, if both X_1 and X_2 describe situations in which all the players choose *rationally*, then why should the players believe that outcomes in X_1 are more plausible than outcomes in X_2? Researchers interested in the epistemic foundations of iterated removal of weakly dominated strategies have discussed this issue (Cubitt and Sugden 2003; Samuelson 1992). For instance, Cubitt and Sugden impose a "privacy of tie-breaking" property, which says that a player cannot *know* that her opponent will not pick an option that is classified as "choice-worthy" (Cubitt and Sugden 2014, p. 8).[9] In our setting, this issue arises because, in general, for events $E_1, \ldots, E_k$:

$$Min_{\preceq}(E_1 \cup E_2 \cup \cdots \cup E_k) \neq Min_{\preceq}(E_1) \cup Min_{\preceq}(E_2) \cup \cdots \cup Min_{\preceq}(E_k).$$

Returning to our example in the previous paragraph, the gray shaded regions identify the most plausible states in X_1 and X_2. We have that $Min_{\preceq}(X_1 \cup X_2) = A \neq Min_{\preceq}(X_1) \cup Min_{\preceq}(X_2) = A \cup E$. The generalization of conservative upgrade that incorporates a constraint analogous to Cubitt and Sugden's privacy of tie-breaking property should result in the following plausibility ordering:

$$A \cup E \prec B \prec C \cup D \prec F \cup G.$$

The formal definition is:

Definition 6.2.6 (Generalized Conservative Upgrade). Let $\mathcal{M} = \langle W, \preceq, \sigma \rangle$ be a plausibility model and $\{E_1, \ldots, E_k\}$ a set of events. Define $\mathcal{M}^{\uparrow\{E_1,\ldots,E_k\}} = \langle W^{\uparrow\{E_1,\ldots,E_k\}}, \preceq^{\uparrow\{E_1,\ldots,E_k\}}, \sigma^{\uparrow\{E_1,\ldots,E_k\}} \rangle$ as follows: $W^{\uparrow\{E_1,\ldots,E_k\}} = W$, $\sigma^{\uparrow\{E_1,\ldots,E_k\}} = \sigma$ and, if $B = Min_{\preceq}(E_1) \cup Min_{\preceq}(E_2) \cup \cdots \cup Min_{\preceq}(E_k)$, then

1. if $v \in B$, then $v \preceq^{\uparrow\{E_1,\ldots,E_k\}} x$ for all $x \in W$; and
2. for all $x, y \in W - B$, $x \preceq^{\uparrow\{E_1,\ldots,E_k\}} y$ iff $x \preceq y$.

Remark 6.2.7 (Suspending Judgement). A generalized conservative upgrade with $\{E, \overline{E}\}$, where $\overline{E}$ is the complement of E, can be interpreted as a *suspension of judgement* regarding E (cf. Holliday (2009) for a discussion). We do not offer an extended discussion of belief suspension here, but we suggest that a natural response

[9] Rabinovich takes this even further and argues that from the principle of indifference, players must assign equal probability to all choice-worthy options (Rabinowicz 1992).

is to learning that there are more than one chose-worthy action for the players is to *suspend judgement* about which options the relevant players will *pick*.

A generalized conservative upgrade of $\{E_1, \ldots, E_k\}$ "flattens" out the players' beliefs relative to this set of events. After the upgrade, the player will consider each of the E_i equally plausible. But this means that, if w is a most plausible E_i-world and v is a most plausible E_j-world, the player forgets whatever reason she had for considering state w more plausible than v (or vice versa). This suggests a generalization of *radical upgrade* , where the players remember their earlier reasons for considering some states more plausible than others. The idea is to update with a set of events as in Definition 6.2.6, but to maintain the original ordering within the union of the most plausible E_i-worlds.

Definition 6.2.8 (Generalized Radical Upgrade). Let $\mathcal{M} = \langle W, \preceq, \sigma \rangle$ be a plausibility model and $\{E_1, \ldots, E_k\}$ a set of events. Define $\mathcal{M}^{\Uparrow\{E_1,\ldots,E_k\}} = \langle W^{\Uparrow\{E_1,\ldots,E_k\}}, \preceq^{\Uparrow\{E_1,\ldots,E_k\}}, \sigma^{\Uparrow\{E_1,\ldots,E_k\}} \rangle$ as follows: $W^{\Uparrow\{E_1,\ldots,E_k\}} = W$, $\sigma^{\Uparrow\{E_1,\ldots,E_k\}} = \sigma$ and, if $B = Min_{\preceq}(E_1) \cup Min_{\preceq}(E_2) \cup \cdots \cup Min_{\preceq}(E_k)$, then

1. for all $v \in B$, $v \preceq^{\Uparrow\{E_1,\ldots,E_k\}} x$ for all $x \in W - B$;
2. for all $x, y \in B$, $x \preceq^{\Uparrow\{E_1,\ldots,E_k\}} y$ iff $x \preceq y$; and
3. for all $x, y \in W - B$, $x \preceq^{\Uparrow\{E_1,\ldots,E_k\}} y$ iff $x \preceq y$.

Applying this definition to the running example in this section results in the plausbility ordering:

$$A \prec E \prec B \prec C \cup D \prec F \cup G.$$

We will see other examples of the transformations defined above in the next section. These transformations can be logically analyzed using standard techniques from dynamic epistemic/doxastic logic literature (e.g., the "reduction axiom method").

6.3 Rational Deliberation via Iterated Belief Updates

In this section, we use the ideas developed in Sect. 6.2 to formally define our model of deliberation in games. The idea is that a player's "rational response" to a given categorization is to transform the current informational context using one of the transformations from the Sect. 6.2.2. To make this precise, we need to *describe* a categorization.

Definition 6.3.1 (Language for a Game). Let $G = \langle N, \{S_i\}_{i \in N}, \{u_i\}_{i \in N} \rangle$ be a finite strategic game. Without loss of generality, assume that each of the S_i are disjoint, and let $\mathsf{At}_G = \{P_a^i \mid a \in S_i,\ i \in N\}$ be a set of atomic formulas (one for each $a \in S_i$). The propositional language for G, denoted $\mathcal{L}_G$, is the smallest set of

formulas containing At_G and closed under the boolean connectives $\neg$ and $\wedge$. The other boolean connectives ($\vee$, $\rightarrow$, $\leftrightarrow$) are defined as usual.

Formulas of $\mathcal{L}_G$ are intended to describe possible outcomes of the game. Given a game model $\mathcal{M}_G$, the formulas $\varphi \in \mathcal{L}_G$ is can be associated with subsets of the set of states in the usual way:

Definition 6.3.2 (Interpretation of $\mathcal{L}_G$). Let G be a strategic game, $\mathcal{M}_G = \langle W, \preceq, \sigma \rangle$ an informational context of G and $\mathcal{L}_G$ a propositional language for G. We define a map $[\![\cdot]\!]_{\mathcal{M}_G} : \mathcal{L}_G \rightarrow \wp(W)$ by induction on the structure of $\mathcal{L}_G$ as follows: $[\![P_a^i]\!]_{\mathcal{M}_G} = \{w \mid \sigma_i(w) = a\}$, $[\![\neg\varphi]\!]_{\mathcal{M}_G} = W - [\![\varphi]\!]_{\mathcal{M}_G}$ and $[\![\varphi \wedge \psi]\!]_{\mathcal{M}_G} = [\![\varphi]\!]_{\mathcal{M}_G} \cap [\![\psi]\!]_{\mathcal{M}_G}$.

Let X and Y be two sets of propositions; we define $X \wedge Y := \{\varphi \wedge \psi \mid \varphi \in X, \psi \in Y\}$

Definition 6.3.3 (Describing a categorization). Let G be a finite game and $\mathcal{M}_G$ an informational context of G. Given a categorization $\mathbf{S}(\mathcal{M}_G)$, let $Do(\mathbf{S}(\mathcal{M}_G))$ denote the set of formulas that *describe* $\mathbf{S}$. This set is defined as follows. For each $i \in N$, let:

$$Do_i(\mathbf{S}_i(\mathcal{M}_G)) = \{P_a^i \mid a \in \mathbf{S}_i^+(\mathcal{M}_G)\} \cup \{\neg P_b^i \mid b \in \mathbf{S}_i^-(\mathcal{M}_G)\}.$$

Then, define $Do(\mathbf{S}(\mathcal{M}_G)) = Do_i(\mathbf{S}_i(\mathcal{M}_G)) \bigwedge Do_2(\mathbf{S}_2(\mathcal{M}_G)) \cdots \bigwedge Do_n(\mathbf{S}_n(\mathcal{M}_G))$.

The general project is to understand the interaction between types of categorizations (e.g., choice rules) and types of model transformations (representing the rational deliberation process). One key question, is whether (and under what conditions) a deliberation process *stabilizes*? There are a number of ways to make precise what it means to stabilize (see Baltag and Smets (2009) for a discussion).

Definition 6.3.4 (Stable in Beliefs). Suppose that $\mathcal{M} = \langle W, \preceq, \sigma \rangle$ and $\mathcal{M}' = \langle W, \preceq', \sigma' \rangle$ are two plausibility models based on the same set of states.[10] We say that $\mathcal{M}$ and $\mathcal{M}'$ are **stable with respect to the players' beliefs** if the set of propositions that are believed in $\mathcal{M}$ is the same as those believed in $\mathcal{M}'$. Equivalently, $\mathcal{M}$ and $\mathcal{M}'$ are stable with respect to beliefs provided $Min_{\preceq}(W) = Min_{\preceq'}(W)$. We write $\mathcal{M} \equiv_B \mathcal{M}'$ when $\mathcal{M}$ and $\mathcal{M}'$ are stable with respect to beliefs.

In this paper, it is enough to define stabilization in terms of the players' simple beliefs because, during the deliberation process, we incorporate only information about what the players are going to do (as opposed to higher-order information[11]). We are now ready to formally define a "deliberation sequence":

[10] So, we assume that the models agree about which outcomes of the game have not been ruled out.

[11] An interesting extension would be to start with a multiagent belief model and allow players to incorporate information not only about which options are "choice-worthy", but also about which beliefs their opponents may have. We leave this extension for future work and focus on setting up the basic framework here.

Definition 6.3.5 (Upgrade Sequence). Given a game G and an informational context $\mathcal{M}_G$, an upgrade sequence of type τ, induced by $\mathcal{M}_G$ is a sequence of plausibility models $(\mathcal{M}_m)_{m\in\mathbb{N}}$ defined as follows:

$$\mathcal{M}_0 = \mathcal{M}_G \qquad \mathcal{M}_{m+1} = \tau(\mathcal{M}_m, Do(\mathcal{M}_m)).$$

An upgrade sequence **stabilizes** if there is an $n \geq 0$ such that $\mathcal{M}_n \equiv_B \mathcal{M}_{n+1}$. The next section has a number of examples of upgrade sequences, some that stabilize and others that do not stabilize.

In the remainder of this section, we discuss a number of abstract principles that any upgrade sequence should satisfy. To state these properties, we need some notation. Let U be a fixed set of states and G a fixed strategic game. We restrict attention to transformations between models of G whose states come from the same set of states U. Let $\mathbb{M}_G$ be the set of all such plausibility models. A model transformation is a function that maps a model of G and a finite set of formulas of $\mathcal{L}_G$ to a model in $\mathbb{M}_G$:

$$\tau : \mathbb{M}_G \times \wp_{<\omega}(\mathcal{L}_G) \to \mathbb{M}_G,$$

where $\wp_{<\omega}(\mathcal{L}_G)$ is the set of finite subsets of $\mathcal{L}_G$. Of course, not all functions τ make sense, given that we intend τ to model belief changes as the players deliberate about what to do. The first set of principles ensure that the categorizations are "sensitive" to the players' beliefs and that the players respond to the categorizations in the appropriate way. Suppose that $\mathcal{X} = \{\varphi_1, \ldots, \varphi_k\}$ is a finite set of $\mathcal{L}_G$ formulas and $\mathcal{M} \in \mathbb{M}_G$.

A1 The operation τ depends only on the truth set of the formulas: If, for each $i = 1, \ldots, k$, $[\![\varphi_i]\!]_{\mathcal{M}} = [\![\psi_i]\!]_{\mathcal{M}}$, then $\tau(\mathcal{M}, \mathcal{X}) = \tau(\mathcal{M}, \{\psi_1, \ldots, \psi_n\})$.

A2 The operation τ is idempotent[12] in the language $\mathcal{L}_G$: $\tau(\mathcal{M}, \mathcal{X}) = \tau(\mathcal{M}^{\tau(\mathcal{X})}, \mathcal{X})$.

Property **A1** says that the belief transformations depend only on the propositions expressed by a formula by treating equivalent formulas the same way. The second property **A2** says that receiving the exact same information twice does not have any effect on the players' beliefs. These are natural properties that are satisfied by any belief-change policy. Certainly, there may be other properties that one may want to impose (for example, variants of the AGM postulates Alchourrón et al. 1985). We leave a discussion of additional principles for another paper. The next two properties ensure that the transformation responds "properly" to a categorization.

A3 For all models $\mathcal{M}, \mathcal{M}' \in \mathbb{M}_G$ and categorizations $\mathbf{S}$, if $\mathcal{M} \equiv_B \mathcal{M}'$, then $\mathbf{S}(\mathcal{M}) = \mathbf{S}(\mathcal{M}')$.

A4 For all models $\mathcal{M}, \mathcal{M}' \in \mathbb{M}_G$, $\tau(\mathcal{M}, Do(\mathbf{S}(\mathcal{M})))$ incorporates $\mathbf{S}(\mathcal{M})$.

[12]Here, it is crucial that the language $\mathcal{L}_G$ does not contain any modalities.

Property **A3** guarantees that the categorization depends only on the players' beliefs. Property **A4** ensures the players are responding to the categorizations in the right way. The properties **A1**, **A2**, **A3** and **A4** are the minimal set of principles that an upgrade sequence must satisfy in order to serve as a model of deliberation in games. We conclude this section by discussing conditions that guarantee that an upgrade sequence will stabilize.

There are two main reasons why an upgrade sequence would stabilize. The first is due to the properties of the transformation (for example, it is clear that upgrade streams with public announcements always stabilize). The second is because the choice rule satisfies a monotonicity property so that, eventually, the categorizations stabilize, and so, there is no new information to change the plausibility ordering. One way to guarantee that upgrade sequences stabilize is to assume that the categorizations satisfy a monotonicity property.

Mon$^-$ For any upgrade sequence $(\mathcal{M}_n)_{n\in\mathbb{N}}$, for all $n \geq 0$, for all players $i \in N$, $\mathbf{S}_i^-(\mathcal{M}_n) \subseteq \mathbf{S}_i^-(\mathcal{M}_{n+1})$.

Mon$^+$ Either for all models $\mathcal{M}_G$, $\mathbf{S}_i^+(\mathcal{M}_G) = S_i - \mathbf{S}_i^-(\mathcal{M}_G)$ or for any upgrade sequence $(\mathcal{M}_n)_{n\in\mathbb{N}}$, for all $n \geq 0$, for all players $i \in N$, $\mathbf{S}_i^+(\mathcal{M}_n) \subseteq \mathbf{S}_i^+(\mathcal{M}_{n+1})$.

Property **Mon**$^-$ means that once an option for a player is classified as "not rationally permissible", it cannot, at a later stage of the deliberation process, drop this classification. Property **Mon**$^+$ says that either the rationally permissible options satisfy the same monotonicity property or they are completely determined by the set of rationally impermissible options.

Theorem 6.3.6. *Suppose that G is a finite strategic game and that all of the above properties are satisfied. Then, every upgrade sequence* $(\mathcal{M}_n)_{n\in\mathbb{N}}$ *for G stabilizes.*

Proof. Let $G = \langle N, \{S_i\}_{i\in N}, \{u_i\}_{i\in N}\rangle$ be a finite strategic game. By properties **Mon**$^-$ and **Mon**$^+$ we have either for all upgrade streams $(\mathcal{M}_n)_{n\in\mathbb{N}}$ and players $i \in N$,

1. $\mathbf{S}_i^-(\mathcal{M}_0) \subseteq \mathbf{S}_i^-(\mathcal{M}_1) \subseteq \cdots \mathbf{S}_i^-(\mathcal{M}_n) \subseteq \cdots$ is an infinitely increasing sequence of subsets of S_i and $\mathbf{S}_i^+(\mathcal{M}_0) \supseteq \mathbf{S}_i^+(\mathcal{M}_1) \supseteq \cdots \mathbf{S}_i^+(\mathcal{M}_n) \supseteq \cdots$ is an infinite decreasing sequence of subsets of S_i; or
2. Both,

$$\mathbf{S}_i^-(\mathcal{M}_0) \subseteq \mathbf{S}_i^-(\mathcal{M}_1) \subseteq \cdots \mathbf{S}_i^-(\mathcal{M}_n) \subseteq \cdots$$

and

$$\mathbf{S}_i^+(\mathcal{M}_0) \subseteq \mathbf{S}_i^+(\mathcal{M}_1) \subseteq \cdots \mathbf{S}_i^+(\mathcal{M}_n) \subseteq \cdots$$

are infinite increasing sequences of subsets of S_i.

Since each S_i is assumed to be finite, for each player i, there is an n_i such that $\mathbf{S}_i^-(\mathcal{M}_{n_i}) = \mathbf{S}_i^-(\mathcal{M}_{n_i+i})$ and $\mathbf{S}_i^+(\mathcal{M}_{n_i}) = \mathbf{S}_i^+(\mathcal{M}_{n_i+i})$. Let m be the maximum of $\{n_i \mid i \in N\}$. Then, we have $\mathbf{S}(\mathcal{M}_m) = \mathbf{S}(\mathcal{M}_{m+1})$. All that remains is to show that

for all $x > m$, $\mathcal{M}_x = \tau(\mathcal{M}_x)$. This follows by an easy induction on x. The key calculation is: for each $x \in \mathbb{N}$, let $\mathcal{D}_x$ be the appropriate description of $\mathbf{S}(\mathcal{M}_x)$.

$$\begin{aligned}\mathcal{M}_{m+2} = \tau(\mathcal{M}_{m+1}, \mathcal{D}_{m+1}) &= \tau(\mathcal{M}_m^{\tau(\mathcal{D}_m)}, \mathcal{D}_{m+1}) \\ &= \tau(\mathcal{M}_m^{\tau(\mathcal{D}_m)}, \mathcal{D}_m) \text{ (since } \mathbf{S}(\mathcal{M}_m) = \mathbf{S}(\mathcal{M}_{m+1})) \\ &= \tau(\mathcal{M}_m, \mathcal{D}_m) = \mathcal{M}_{m+1}\end{aligned}$$

This concludes the proof. QED

A number of researchers have noticed that monotonicity of the choice rule is important for an epistemic analysis of games (see Apt and Zvesper (2010b) for a discussion). An immediate corollary of Theorem 6.3.6 is:

Corollary 6.3.7. *If the categorization method is strict dominance, then any upgrade sequence of type τ stabilizes, where τ is any of the transformations discussed in this paper (e.g., public announcement, (generalized) radical upgrade and (generalized) conservative upgrade).*

This is related to van Benthem's iterated "soft" announcements of rationality (van Benthem 2007) and Apt and Zvesper's results about stabilization of beliefs in games (Apt and Zvesper 2010a).

6.4 Case Study: Iterated Weak Dominance

Larry Samuelson (1992) points out an explicit puzzle surrounding the epistemic foundations of iterated removal of weakly dominated strategies (IEWDS) – also known as the IA solution. He shows (among other things) that there is no epistemic model of the following game with at least one state satisfying "common knowledge of admissibility" (i.e., a state in which there is common knowledge that the players do not play a strategy that is weakly dominated).

		Bob	
		L	R
Ann	u	$1,1$	$1,0$
	d	$1,0$	$0,1$

In the above game, d is weakly dominated by u for Ann. If Bob believes that Ann is rational (in the sense that she will not choose a weakly dominated strategy), then he can conclude that u is more plausible than d. In the smaller game, action R is now strictly dominated by L for Bob. If Ann believes that Bob is rational and that Bob knows that she is rational (and thus, d is rationally impermissible), then she can conclude that L is more plausible than R. Assuming that the above reasoning is transparent to both Ann and Bob, it is common knowledge that Ann will play u and Bob will play L. But now, what is the reason for Bob to rule out the possibility that

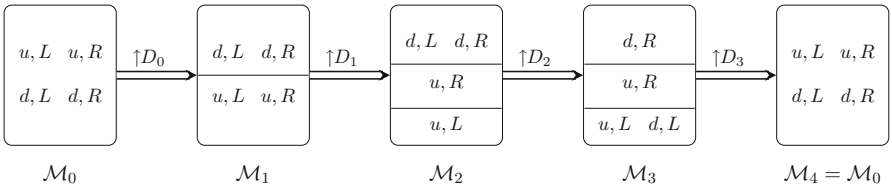
Ann will play d? He believes that Ann believes that he is going to play L, and both u and d are best responses to L.

The general framework introduced in Sects. 6.2 and 6.3 offers a new, dynamic perspective on Samuelson's analysis, as well as on reasoning with weak dominance more generally. Note that we are not providing an *alternative* epistemic characterization of IEWDS. Both Brandenburger et al. (2008) and Halpern and Pass (2009) have convincing results here. Our goal is to use this solution concept to illustrate our general approach.

Generalized Conservative Upgrade with Weak Dominance Dynamically, Samuelson's analysis of the above game corresponds to non-stabilization of an upgrade sequence. The players are not able to reason their way to stable, common belief in admissibility. To capture this intuition, in light of Theorem 6.3.6, we need to work with a non-monotonic categorization. Before stating the observation formally, we need one more definition. A **full model** of a game G is one in which all outcomes of the game are in the model (i.e., for any profile $\vec{s}$, there is a state w satisfying $\sigma(w) = \vec{s}$) and the states are all equally plausible.

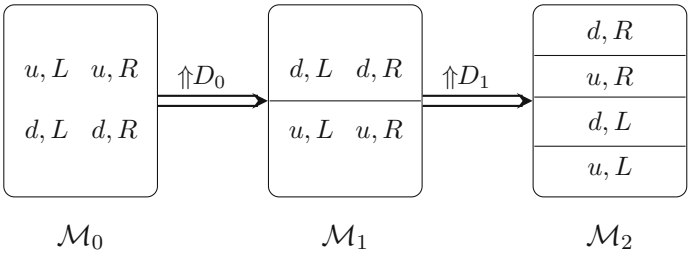
Observation 6.4.7 *Starting with the initial full model of the above game, the conservative upgrade sequence for conservative upgrade and weak dominance does not stabilize.*

The proof of this Observation is provided by the following looping stream of conservative upgrades:



where for $i = 1, 2, 3, 4$, $D_i = Do(\mathbf{WD}(\mathcal{M}_i))$. Intuitively, from $\mathcal{M}_0$ to $\mathcal{M}_2$ the players have reasons to exclude d and R, leading them to commonly believe that u, L is played. At that stage, however, d is admissible for Ann, canceling the players' reason for ruling out this strategy. The rational response is, thus, to suspend judgment on d, leading to $\mathcal{M}_3$. In this new model, the agents are similarly led to suspend judgment on not playing R, bringing them back to $\mathcal{M}_0$. This process loops forever; the agents' reasoning does not stabilize.

Generalize Radical Upgrade with Weak Dominance Generalized radical upgrade stabilizes plain beliefs even for non-monotonic choice rules such as weak dominance. Consider, again, Samuelson's game given above. Starting with the full model of this game, the upgrade stream stabilizes on a model with the (common) belief that all the players will play the IEWDS outcome.



where D_0 and D_1 are as above. Intuitively, what happens is the following: Just as with conservative upgrade, $\mathcal{M}_0$ and $\mathcal{M}_1$, respectively, give the agents reasons to believe that Ann will not play d, and that Bob will not play R. This leads to $\mathcal{M}_2$, where, like before, d is admissible given that Ann believes that Bob will play L. Radical upgrade, however, does not allow this fact to override her reason for not playing d: her rational response is to rank u, L and d, L above all other possible outcomes, but to keep the relative ordering of these two, reflecting the fact that she previously ruled out u.

Stabilization of radical upgrade puts Samuelson's observation into perspective. Such an upgrade forces the agents to remember the reasons they had earlier in the deliberation. Previous reasons constrain the domain of permissibility at later stages in the deliberation process. What is permissible for Ann at $\mathcal{M}_2$ depends on the deliberation process that led to this model, and, in particular, on the existence of an (earlier) reason not to play d. This was not the case for conservative upgrade. Reasons at each stage were evaluated de novo, without reference to the reasoning history. This is what led the upgrade sequence for Samuelson's game into looping, to the "paradox" of admissibility. We leave open for discussion whether this constitutes an argument to the effect that players "should" keep track of their reasons while reasoning to a specific informational context. For now, we content ourselves with the observation that there is a tight connection, on the one hand, between remembering one's reasons and stabilization of reasoning under admissibility and, on the other hand, between letting new reasons override previous ones and the possibility of never-ending reasoning chains.

6.5 Concluding Remarks

A general theory of rational deliberation for game and decision theory is a big topic, and, thus, it is beyond the scope of this article to discuss the many different aspects and competing perspectives on such a theory. The reader is referred to Brian Skyrms' (1990, Chap.7) for a broader discussion. The main contribution of this paper is to lay the foundation for a formal theory of deliberation in games, based on recent work on dynamic logics of knowledge and belief. We focused on one specific question: What type of process can be used to *generate* a game model?

The most pressing philosophical issue concerns the role that a theory of deliberation plays in rational choice theory (cf. Levi 1993; Rabinowicz 2002; Schick 1979; Arntzenius 2008). On the technical side, throughout the paper, we worked with (logical) models of *all out* attitudes, leaving aside probabilistic, graded beliefs, even though they are arguably the most widely used in the current literature on epistemic foundations of game theory. It is an important, and non-trivial, task to transpose the dynamic perspective on informational contexts that we advocate here to such probabilistic models. We leave that for future work.

Finally, we should stress that the dynamic perspective on informational contexts is a natural complement, and not an alternative, to existing epistemic characterizations of solution concepts (van Benthem et al. 2011). Epistemic characterizations of solution concepts offer rich insights into the consequences of taking the informational contexts of strategic interaction seriously. What we proposed here is a first step towards understanding how and why such a context might arise.

References

Alchourrón, C.E., Gärdenfors, P., Makinson, D.: On the logic of theory change: partial meet contraction and revision functions. J. Symb. Log. **50**, 510–530 (1985)

Anglberger, A.J.J., Gratzl, N., Roy, O.: Obligation, free choice and the logic of weakest permission. Rev. Symb. Log. **8**(4), 807–827 (2015)

Apt, K., Zvesper, J.: Public announcements in strategic games with arbitrary strategy sets. In: Proceedings of LOFT 2010, Toulouse, France (2010a)

Apt, K., Zvesper, J.: The role of monotonicity in the epistemic analysis of strategic games. Games **1**(4), 381–394 (2010b)

Arntzenius, F.: No regrest, or: edith piaf revamps decision theory. Erkenntnis **68**, 277–297 (2008)

Baltag, A., Smets, S.: Group belief dynamics under iterated revision: fixed points and cycles of joint upgrades. In: Proceedings of Theoretical Aspects of Rationality and Knowledge, Stanford (2009)

Baltag, A., Smets, S., Zvesper, J.: Keep 'hoping' for rationality: a solution to the backwards induction paradox. Synthese **169**, 301–333 (2009)

Board, O.: Dynamic interactive epistemology. Games Econ. Behav. **49**, 49–80 (2004)

Boutilier, C.: Conditional logics for default reasoning and belief revision. PhD thesis, University of Toronto (1992)

Brandenburger, A., Friedenberg, A., Keisler, H.J.: Admissibility in games. Econometrica **76**, 307–352 (2008)

Cubitt, R., Sugden, R.: Common knowledge, salience and convention: a reconstruction of David Lewis' game theory. Econ. Philos. **19**(2), 175–210 (2003)

Cubitt, R., Sugden, R.: The reasoning-based expected utility procedure. Games Econ. Behav. **71**(2), 328–338 (2011)

Cubitt, R., Sugden, R.: Common reasoning in games: a Lewisian analysis of common knowledge of rationality. Econ. Philos. **30**(3), 285–329 (2014)

Dekel, E., Siniscalchi, M.: Epistemic game theory. In: Peyton Young, H., Shmuel, Z. (eds.) Handbook of Game Theory with Economic Applications, vol. 4, pp. 619–702. Elsevier, Amsterdam (2015)

Gerbrandy, J.: Bisimulations on planet Kripke. PhD thesis, University of Amsterdam (1999)

Halpern, J., Pass, R.: A logical characterization of iterated admissibility. In: Heifetz, A. (ed.) Proceedings of the Twelfth Conference on Theoretical Aspects of Rationality and Knoweldge, Stanford, pp. 146–155 (2009)

Harsanyi, J.: The tracing procedure: a Bayesian approach to defining a solution for *n*-person noncooperative games. Int. J. Game Theory **4**, 61–94 (1975)
Holliday, W.: Trust and the dynamics of testimony. In: Logic and interaction rationality: Seminar's yearbook 2009, pp. 147–178. ILLC technical reports (2009)
Lamarre, P., Shoham, Y.: Knowledge, certainty, belief and conditionalisation. In: Proceedings of the International Conference on Knoweldge Representation and Reasoning, pp. 415–424 (1994)
Levi, I.: Rationality, prediction, and autonomous choice. Can. J. Philos. **23**(1), 339–363 (1993)
Lewis, D.: Convention. Harvard University Press, Cambridge (1969)
Lewis, D.: Counterfactuals. Blackwell, Oxford (1973)
Leyton-Brown, K., Shoham, Y.: Essentials of Game Theory: A Concise Multidisciplinary Introduction. Morgan & Claypool, San Rafael (2008)
Pacuit, E.: Dynamic epistemic logic II: logics of informaiton change. Philos. Compass **8**(9), 815–833 (2013)
Pacuit, E.: Dynamic models of rational deliberation in games. In: van Benthem, J., Ghosh, S., Verbrugge, R. (eds.) Models of Strategic Reasoning: Logics, Games and Communities. Springer (2015)
Pacuit, E., Roy, O.: Epistemic foundations of game theory. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy (2015)
Perea, A.: Epistemic Game Theory: Reasoning and Choice. Cambridge University Press, New York (2012)
Peterson, M.: An Introduction to Decision Theory. Cambridge University Press, Cambridge/New York (2009)
Plaza, J.: Logics of public communications. In: Emrich, M.L., Pfeifer, M.S., Hadzikadic, M., Ras, Z.W. (eds.) Proceedings, 4th International Symposium on Methodologies for Intelligent Systems, pp. 201–216 (1989)
Rabinowicz, W.: Tortous labyrinth: noncooperative normal-form games between hyperrational players. In: Bicchieri, C., Chiara, M.L.D. (eds.) Knowledge, Belief and Strategic Interaction. Cambridge University Press, Cambridge/New York, pp. 107–125 (1992)
Rabinowicz, W.: Does practical deliberation crowd out self-prediction. Erkenntnis **57**, 91–122 (2002)
Rott, H.: Shifting priorities: simple representations for 27 itereated theory change operators. In: Lagerlund, H., Lindström, S., Sliwinski, R. (eds.) Modality Matters: Twenty-Five Essays in Honour of Krister Segerberg. Uppsala Philosophical Studies, vol. 53, pp. 359–384 (2006)
Roy, O., Anglberger, A.J.J., Gratzl, N.: The logic of best action from a deontic perspective. In: Baltag, A., Smets, S. (eds.) Johan FAK van Benthem on Logical and Informational Dynamics. Springer (2014)
Samuelson, L.: Dominated strategies and common knowledge. Game Econ. Behav. **4**, 284–313 (1992)
Schick, F.: Self-knowledge, uncertainty and choice. Br. J. Philos. Sci. **30**(3), 235–252 (1979)
Skyrms, B.: The Dynamics of Rational Deliberation. Harvard University Press, Cambridge (1990)
Ullmann-Margalit, E., Morgenbesser, S.: Picking and choosing. Soc. Res. **44**, 757–785 (1977)
van Benthem, J.: Dynamic logic for belief revision. J. Appl. Non-class. Log. **14**(2), 129–155 (2004)
van Benthem, J.: Rational dynamics and epistemic logic in games. Int. Game Theory Rev. **9**(1), 13–45 (2007)
van Benthem, J.: Logical Dynamics of Information and Interaction. Cambridge University Press (2010)
van Benthem, J.: Logic in Games. MIT, Cambridge/London (2014)
van Benthem, J., Gheerbrant, A.: Game solution, epistemic dynamics and fixed-point logics. Fund. Inf. **100**, 1–23 (2010)
van Benthem, J., Pacuit, E., Roy, O.: Towards a theory of play: a logical perspective on games and interaction. Games **2**(1), 52–86 (2011)

Chapter 7
Relevant Alternatives in Epistemology and Logic

Peter Hawke

Abstract The goal of the current paper is to provide an introduction to and survey of the diverse landscape of *relevant alternatives theories of knowledge*. Emphasis is placed throughout both on the abstractness of the relevant alternatives approach and its amenability to formalization through logical techniques. We present some of the important motivations for adopting the relevant alternatives approach; briefly explore the connections and contrasts between the relevant alternatives approach and related developments in logic, epistemology and philosophy of science; provide a schema for classifying and studying relevant alternatives theories at different levels of abstraction; and present a sample of relevant alternatives theories (contrasting what we call question-first and topic-first theories) that tie our discussion to ongoing debates in the philosophical literature, as well as showcasing techniques for formalizing some of the important positions in these debates.

Keywords Relevant alternatives theory • Epistemic relevance • Epistemic closure • Epistemic logic • Questions • Subject matter

7.1 Introduction

The aim of the current paper is to introduce the reader to the relevant alternatives (RA) approach to the theory of knowledge and provide some indication of the complex landscape such theories inhabit.

One important theme that we emphasize throughout is the breadth and versatility of the RA approach – at least in the very general form we expound and develop it here. Indeed, the diversity of the existing theories of knowledge that fall under the RA banner – many of which we will meet in this paper, notably in Sects. 7.2.8 and 7.4 – bears testimony to the *abstractness* of our basic RA framework.

Another important theme is that RA theory, in its many guises, is typically amenable to study using precise formal methods. The RA approach is, therefore,

P. Hawke (✉)
Department of Philosophy, Stanford University, Building 90, 450 Serra mall, Stanford, CA 94305-2155, USA
e-mail: phawke@stanford.edu

J. Redmond et al. (eds.), *Epistemology, Knowledge and the Impact of Interaction*, Logic, Epistemology, and the Unity of Science 38, DOI 10.1007/978-3-319-26506-3_7

not only a unifying framework for diverse, nuanced and intriguing philosophical theories of knowledge (encompassing, as should become apparent, a significant bulk of important recent developments for this ancient philosophical inquiry), but is also a notable site for the interaction between epistemology and logic. This interaction extends fruitfully in both directions (Hawke 2016; Holliday 2012, 2014, 2015): logical techniques allow the RA theorist to operate at an unusual level of technical precision when framing rival positions and their consequences, informing the philosophical discussion in a manner that goes beyond mere window-dressing; while RA theory is a source of novel, sophisticated variants of epistemic logic, worthy of detailed logical study in their own right.

It is *not* a goal of the current paper to defend or attack any particular RA theory, or even the RA approach as a whole. Rather: it is to awaken in the reader an interest in RA theory as a venue for both epistemology and logic, illustrate the scope and dimensions of the RA approach and broach interesting questions for the RA theorist.

In the next section, I introduce the reader to the spirit of RA theory and review the motivation for this general approach, citing, for instance, some compelling linguistic considerations and the idea that RA theory captures the 'common man' response to the problem of cartesian skepticism. In addition, we briefly draw out some connections and contrasts between the RA approach and similarly themed discussions in the logic, epistemology and scientific methodology literature. In the third section, I propose a series of basic 'choice points' for the RA theorist. The leading claim here is that any *particular*, 'concrete' theory of knowledge that counts as an RA theory is essentially the product of settling each choice point. Hence, our list of choice points, it is suggested, offers a basic schema for *classification* of RA theories and provides a tool for studying RA theory at different levels of abstraction (where a higher level of abstraction corresponds to leaving more choice points open). We discuss each choice point in turn and briefly mention techniques for formalizing some of the potential paths associated with each choice point that the RA theorist can follow. In the fourth section, I exhibit a number of RA theories, suitably formalized using logical semantics, and classify them according to the schema from Sect. 7.3. I associate these theories with concrete proposals from leading contemporary writers in the philosophical literature – specifically, Jonathan Schaffer and Stephen Yablo – and connect our discussion with important recent philosophical debates. With that, we conclude.

7.2 RA Theory: Its Nature and Motivation

7.2.1 The Slogan

The spirit of RA theory is quickly captured by the following slogan:

> In order for S to know that P, S need only have evidence that rules out all of the *relevant* alternatives to P (that is, S need *not* have evidence that rules out *all* of the alternatives to P).

Intuitively, one can think of an alternative to P as a circumstance that conflicts with P (we deliberately here use 'circumstance' in a vague and intuitive way, neutral between 'proposition', 'claim', 'state of affairs', 'possible world' and so forth). An alternative to the circumstance that Bush won the election is that Gore won the election. An alternative to the circumstance that Alan Turing was born in 1913 is that he was born in 1914 (this alternative to the circumstance in question is in fact not the case), and another is that he was born in 1912 (this alternative is in fact the case).[1]

Also intuitively, one may approach the notion of *ruling out* by way of that of *evidence*: the function of evidence is to *rule out* alternatives. If I were a detective trying to solve a murder case and my hypothesis is that the perpetrator was the butler, evidence that establishes that the maid has an alibi is *significant* evidence as it rules out the alternative that the maid is the murderer. Of course, 'ruled out' has an intuitively *strong* reading (to be contrasted with, say, 'unlikely') that seems befitting of association with the term 'knowledge'.

To say then that knowledge of P involves acquiring evidence good enough to rule out all of the alternatives to P has, to many ears, the air of a platitude. The RA theorist thinks that this saying is only half-right, however: coming to know involves ruling out only *select* alternatives, those that are *(epistemically) relevant*.[2]

Obviously, what 'relevance' comes to is a key concern when judging an RA theory. We say more about relevance later (specifically, Sect. 7.3.4). It is worth immediately stoking some intuitions, however: irrelevant alternatives are ones that are (in context) "far-fetched"; are not to be "taken seriously" when making judgements concerning knowledge; are rightly "ignored", in some sense, when it comes to matters epistemic. A suggestive example: to know that my *left* neighbour's dog is barking, I need to rule out that the barking sounds I hear are not emanating from the direction of my *right* neighbour's house. But it might seem that I do *not* need to rule out the bizarre possibility that my left neighbour's dog has been kidnapped and the kidnappers left behind a recording device to play back the sound of a dog's bark as an elaborate ruse.

It should be emphasized from the outset how little content we initially commit to in our introduction of the notion of "relevance" (indeed, this trend continues as we discuss the basic motivations for the RA approach in the coming sections). *All* that we begin with is the idea that (and some reason to think that) some alternatives are relevant to the evaluation of an agent's knowledge, and others are not. In particular, we do *not* in the initial statement of the RA approach commit to the

[1]For this paper, we set aside the tricky question as to in what sense circumstances that are necessarily the case or not the case (such as Goldbach's conjecture) can be thought to have alternatives – at least for the purposes of inquiry.

[2]Note that in Sect. 7.2.9 we discuss a tradition in the epistemology literature that focuses on a notion of "epistemic relevance" that arises from initial concerns quite distinct from those of the RA theorist. The overlapping terminology is no doubt a potential source for confusion, though hopefully not in the current paper.

idea that relevance is a matter of rationality, irrationality or arationality; and we do not commit to the idea that relevance is a function of context, conversational or otherwise. Our statement of the RA approach leaves these questions open.

7.2.2 Motivating the RA Approach

Why be an RA theorist? In the following sections, we present a number of important motivations (many of which, the reader might note, are related, though still worth separating out). Let us begin with an overview.

- There is *striking linguistic data*, concerning our ordinary usage of epistemic claims, that seems to support RA theory.
- The RA approach provides a unique and compelling reply to *cartesian skepticism*.
- More generally, RA theory provides a universal strategy for dealing with *underdetermination problems*.
- RA theory is suggested by our intuitive reaction to *Goldman-Ginet barn cases*.
- RA theory has theoretical value (for contextualists and others) as an attractive and convenient tool for *measuring epistemic standards*.

7.2.3 Suggestive Linguistic Data

Two kinds of purported linguistic data have been used (in concert) to support the RA approach. First, linguistic data seems to indicate a *fallibilist* aspect to our ordinary knowledge concept. That is, it seems that ordinary agents will sometimes happily attribute knowledge to themselves (or others), but, *if pressed*, will concede that certain possibilities for error are compatible with the available evidence. Second, linguistic data seems to indicate an *infallibilist* aspect to our ordinary knowledge concept. That is, ordinary agents seem uncomfortable to state the conjunction of a knowledge claim with an explicit acknowledgement of live possibilities of error. These points are emphasized by both Dretske (1981) and Lewis (1996), following in the footsteps of Unger (1975) (though it should be noted that Unger (1975), a thoroughgoing defense of skepticism, hardly supports an RA approach).

We may immediately note the tension between the seeming fallibilist and infallibilist tendencies of ordinary knowledge ascription. As we shall see, one alleged advantage of the RA approach is its seeming capacity to resolve this tension.

To illustrate the fallibilist tendency, we (ab)use an influential case. Fred Dretske (1970) is famous for pointing out that, under ordinary circumstances (using ordinary visual evidence), one will seem perfectly happy to say that one knows that the animal one sees at the zoo – in the zebra enclosure – is a zebra, but one will be *less* happy, it seems, to say that one knows that the animal is not a mule painted to appear like a zebra. The ordinary visual evidence does not seem to settle the latter

issue. Also famously, Dretske uses this example as a counter-example to the claim that knowledge is closed under known entailment: one can know P, know that P entails Q, put "two and two together" and yet not know that Q.

The legitimacy and exact diagnosis of this purported linguistic data, and its consequences for the truth of the closure principle, are controversial (Vogel 1990; Luper 2001). For our purposes, however, note that the description of the example may be slightly altered in a telling way, by weakening the proposed judgement concerning the "painted mule" possibility: under ordinary circumstances – it is plausibly suggested – one is happy to say one *knows* the enclosed animal is a zebra, yet, if pressed, will be hesitant to add that one has evidence that *rules out* that the animal is a painted mule. Yet being a painted mule is an alternative to being a zebra. Thus, it appears we have everyday linguistic data to the effect that we are often willing to ascribe knowledge of P, yet will quickly concede the limitations of the available evidence when it comes to ruling out *certain* alternatives to P.[3]

As has been pointed out by critics of the RA approach (Vogel 1990, 1999), this modest reading of our intuitions in the zebra case may support fallibilism in general, but does not support the RA approach in particular. For there are other prominent fallibilist approaches to the theory of knowledge. Consider, for instance, the type of Bayesian that holds that knowledge of P is essentially a matter of not-P being *sufficiently improbable* on the evidence. Such a Bayesian, it seems, is a rival to the RA theorist, seeing no need for evidential support (in its role as constraining the space of possibilities) to be supplemented with an independent notion of "relevance".

This form of Bayesianism, however, does not seem as effective in accounting for the *infallibilist* tendencies in our ordinary knowledge ascriptions. Ordinary speakers, it seems, feel uncomfortable in making or accepting claims along the following lines: "I know that P, though not-P might well be the case"; "I know that P, yet my evidence does not vouchsafe certainty that P"; "I know that P, though not-P remains a live possibility". Lewis sums up the sentiment effectively:

> If you claim that S knows that P, and yet you grant that S cannot eliminate a certain possibility that not-P, it certainly seems as if you have granted that S does not after all know that P. To speak of fallible knowledge, of knowledge despite uneliminated possibilities of error, just *sounds* contradictory Lewis (1996, p.549, his emphasis).

The aforementioned Bayesian approach seems an awkward fit with infallibilism. According to this account, one can know P when the probability bestowed on P by the evidence meets an appropriate threshold. But if this threshold is less than 1, then the Bayesian is committed to the possibility that an agent may know that P and yet not-P has non-zero probability and, so, is compatible with (if unlikely on) the evidence.

[3]Throughout this section, the critical reader may well want to emphasize the use of the word *appearance* here in the absence of a proper empirical investigation of these purported linguistic facts.

The RA theorist, on the other hand, has a trick to play. She can account for our fallibilist tendencies: if P is known and yet an alternative A to P is identified as uneliminated by the evidence, then A, the RA theorist proposes, is (or, at least, *was* in context) an irrelevant alternative to the evaluation of knowledge of P. On the other hand, the RA theorist can account for our infallibilist tendencies. She can essentially agree with the ordinary assessment that "to know is to leave *no* possibility for error". But what is left *implicit* in such a saying, the RA theorist proposes, is that the possibilities being quantified over are only the *relevant* ones in that conversational context.[4]

7.2.4 The RA Strategy Against Skepticism

Consider the following argument for a skeptical conclusion:

P1. To be a handless brain-in-a-vat is an alternative to having hands.
P2. The evidence in my possession is not sufficient to rule out that I am a handless brain-in-a-vat.
P3. In order to know P, one needs to have evidence that rules out all alternatives to P.
C. **Therefore:** I do not know that I have hands.

This argument is valid, and P1 and P2 might strike one as undeniable (to deny them, it might be said, is simply not to correctly appreciate the nature of the brain-in-vat scenario). To resist skepticism, there seems only one way out: deny P3. Of course, this is simply to embrace the RA slogan.

I note that, at least in my experience, something along these lines is a common response from the layman (i.e. non-philosophers) when presented with the threat of skepticism. The reaction, it seems, is to deride the brain-in-vat scenario as far-fetched and otherwise *irrelevant* to our ordinary epistemic concerns. Such a reaction seems particularly apt when a practical application of everyday knowledge is afoot. It is in no way an adequate response to the question "do you know where I left my keys?" to say "no, for I cannot rule out that my senses are being deceived by an evil demon". To the extent that she is willing to take the layman as a competent user of the knowledge concept, the RA theorist finds this reaction telling.[5]

[4] Our ordinary infallibilist tendencies have in fact been used as a weapon in *internal* debates among RA theorists, suggesting the possibility that some versions of RA theory are better suited to account for these tendencies than others. For instance, DeRose (1995) influentially criticizes Dretske's version of RA theory as incorrectly predicting that so-called *abominable conjunctions* – notably "S knows that she has hands and S does not know that she is a handless brain-in-vat" – are felicitous in ordinary conversational contexts.

[5] Of course, this may be taken as further linguistic data, in the spirit of that from Sect. 7.2.1.

7.2.5 RA Theory as a Response to Under-determination Problems

Let us generalize the previous point of motivation. Cartesian skepticism, at least in certain forms, is an instance of a larger class of problems that may be called *under-determination problems*. An under-determination problem has the following form: it is obvious that we know that P, yet, on close inspection, our supposed evidence for P seems just as compatible with some (maybe odd, but logically possible) alternative Q. Another prominent under-determination problem: Humean skepticism, where our sensory evidence of particulars seemingly under-determines the general knowledge we tend to hold upon its basis. In general, under-determination has proven a pressing issue in philosophy of science (Stanford 2009).

An RA theory embodies a universal strategy for dealing with under-determination problems: simply establish that any deviant alternatives are properly classified as irrelevant (whatever this classification comes to).

7.2.6 RA Theory by Way of the Goldman-Ginet Barn Case

Along with the Gettier examples and Dretske's painted mule example, the Goldman-Ginet barn case (Goldman 1976) has been a particularly influential example in the contemporary epistemology literature. Suppose subject S clearly observes what is in fact a (genuine) barn out of her car window, as she drives by. Does she know that it is a barn? Our reaction to this question will depend, the example seems to show, on whether S is driving through a county in which the only objects that look like barns to the casual observer are, in fact, barns (in which case, she does know), or if she is in the unusual situation where there are as many barn facades ("fake barns") around as real barns (in which case, she does not).

What exactly does the barn case teach us? The RA theorist may point out the following: it seems to demonstrate that it is possible that S has *exactly the same evidence* in states s and s' (not to mention the same beliefs), and yet S knows that P in s and does not know that P in s' (where P is true in both s and s'). The difference, the RA theorist will urge us: different alternatives to P are relevant in one case than in the other, and, in particular, S does not have sufficient evidence to rule out an alternative (that the object is a barn facade) that happens to be irrelevant in s and happens to be relevant in s'.[6]

[6]Note that the barn case can be seen to teach a similar lesson to consideration of cartesian skepticism: that one can know something even though one has not ruled out *all* alternatives. However, the barn case potentially teaches us something more: that what counts as a relevant alternative can *vary* with the circumstances: the possibility of fake barns may be properly ignored, by knowledge ascribers, under one set of circumstances, but is not properly ignored in another.

7.2.7 RA Theory and Epistemic Standards

A great number of authors in the recent epistemology literature have defended some version of the idea that the *epistemic standards* that an agent needs to meet in order to know that P can vary from context to context. What is chiefly debated, amongst such authors, is *which* context determines the relevant standards: is it that of the *subject* to whom knowledge is potentially attributed (Stanley 2005), that of the *speaker* who is performing the attribution (Cohen 1988; DeRose 1995; Lewis 1996), or that of an *assessor* potentially different to both speaker and subject (MacFarlane 2005)?

Whichever view one takes, such perspectives on the semantics of knowledge claims are a natural fit with RA theory. For how are we to understand the idea of an *epistemic standard*? A natural suggestion is that a variation of epistemic standards consists in a variation of the amount of alternatives that need to be ruled out: a higher standard involves a larger amount of relevant alternatives.

Thus, it is a short path from accepting that epistemic standards vary by context to an acceptance of some form of RA theory. Arguments for the former may therefore, with the right massaging, be taken as support for the latter.[7]

7.2.8 The History of RA Theory

We have thus far discussed RA theory and its motivation as if it exists in a vacuum. In fact, the list of active and explicit defenders of RA theory in the literature is long and varied: Dretske (1970, 1981), Goldman (1976), Luper (1984), Lewis (1996), Cohen (1988), Heller (1989), Pritchard (2012), Lawlor (2013), and Holliday (2015). As I will explain in a moment, it is reasonable to add to this list Austin (1946) and Nozick (1981). Let us briefly delve into some of the history of support for RA theory.

J.L. Austin is notable for making especially early remarks in the direction of an RA theory. His suggestions are discussed and expanded at length by Lawlor (2013).

Fred Dretske (1970, 1981), however, may be singled out as fully initiating the ongoing discussion of RA theory. Dretske's view, somewhat obliquely presented in his initial paper, is roughly as follows: for agent S to know that (true) P, S must believe P on the basis of a conclusive reason R, where R being conclusive means that: if P had not held, then neither would R have held. Averting to the standard ideas in the literature on the semantics for counter-factual conditionals we can say: knowledge requires that in the nearest worlds to actuality in which P is false, so too is R false. We may say then that for Dretske, roughly, an alternative Q to P is relevant just in case it holds at the nearest worlds in which P is false, and Q is ruled out just in case those worlds are incompatible with the agent's reasons (evidence/information).

[7]For a more careful defence of the 'alternatives' approach to capturing the relevant parameter that shifts across contexts, see Schaffer (2005b).

On Dretske's view then, the relevance of Q is relative to the proposition being considered as an object of knowledge: Q might well be relevant relative to one proposition and irrelevant relative to another.

We might contrast this theory to that of Lewis (1996), another prominent and influential RA theorist: roughly, according to Lewis, relevance is determined by a complex set of rules operating on the conversational context in which knowledge attributions may be made. Thus, relevance is relative to a conversational context common to all propositions, as opposed to a specific proposition evaluated for knowledge: if the context is held fixed, proposition Q is fixed as relevant (or irrelevant, as the case may be), no matter which proposition it is contrasted to as an alternative.

I re-emphasize an important point for the current paper brought out by these observations: on occasion in the literature, the label "relevant alternatives theory" is very closely associated with Dretske's theory in particular (and Lewis, for instance, is classified, in contrast, as a "contextualist"). As should by now be evident, in this paper we use the term "relevant alternatives theory" in a liberal and broad manner that encompasses a wide range of views. Indeed, given its structural similarities to Dretske's view (simply replace talk of "having a conclusive reason" with "having a sensitive belief") we could happily class Nozick's well-known tracking theory of knowledge (and its variations) under the RA banner. As we will see in Sect. 7.4, we can also convincingly fit recent work by Schaffer and Yablo under the RA banner, though again these authors do not tend to self-describe their views with this label. In our view, the unifying generality and abstractness of the RA approach is part of its appeal as an object of study.

7.2.9 Connections and Contrasts: Relevance Logic, "Epistemic Relevance" and Scientific Methodology

We close this section with some brief discussion of the potential connections and contrasts between the RA approach and other salient developments in the epistemology and logic literature. Our aim is to achieve some sense of the theoretical promise of the RA approach (insofar as it can integrated and unified with similarly motivated concerns in other strands of the literature) while also being sure to *distinguish* the concerns of the RA theorist from sometimes only superficially similar issues.

(For readers keen to immediately dig into more nitty-gritty features of the RA approach, note that this section may be skipped without any significant break in the flow of the paper).

7.2.9.1 Relevance Logic

Begin with the well-developed field of *relevance logic* (Anderson and Belnap 1975; Burgess 2009; Mares 1998). In brief, what animates this area of logic is a desire to

build (technically and philosophically sound) logics that avoid endorsing so-called "fallacies of relevance" as valid. In particular, the relevance logician is concerned to avoid two counter-intuitive results of classical logic: that any sentence is a valid consequence of contradictory premises, and that a necessary truth is a valid consequence of any set of premises whatsoever. The difficulty with these results, the relevance logician claims, becomes immediately evident when we consider cases where the premises and conclusion are *irrelevant to each other* insofar as they concern disjoint *subject matter*: it does not follow from the claim that the moon is both made of green cheese and not that Barack Obama is president of the USA (nor, for that matter, does it follow from $2+2=5$). Further, it does not follow from the fact that Berlin is the capital of Germany that either it is raining in London or it is not (nor, for that matter, that $2+2=4$).

In sum then (though we place ourselves at risk of over-simplifying), the relevance logician has two concerns: (i) to offer an account of when one proposition is "relevant" to another (which, for all we have said, appears to amount to accounting for what it means to say that two propositions overlap in subject matter) and (ii) an integration of this account into a logical system, to the effect that only *relevant* conclusions are valid consequences of a set of premises. The concerns of the relevance logician and a RA theorist overlap, therefore, to the extent that (i) and (ii) are pertinent to the RA theorist in question.

Is (i) pertinent to an RA theorist? This will depend on whether the RA theorist and relevance logician mean the same thing by "relevance". Since they are motivated by different starting points (which alternatives can be "properly ignored" when evaluating knowledge claims versus what intuitively follows from what) there is no guarantee that there will in general be a convergence here. Indeed, there is a quick argument that "relevance" as deployed by a standard RA theorist must have a different sense (or at least application) than that deployed by the relevance logician. For: suppose we follow the standard line (we return to this in Sect. 7.3.5) and say that proposition A is an alternative to P just in case P entails $\neg A$. Now, the RA theorist wishes to draw a *distinction* between relevant and irrelevant alternatives to P. But this distinction seems to rely on the claim that *both* kinds of alternative are logically related to P. *Both* are "relevant", therefore, in the sense of the relevance logician.

Nevertheless, we discuss in Sect. 7.4 *topic-first RA theorists* that attempt to account for "relevance" in terms of subject matter. For such RA theorists, agreement on considerations of relevance might be sturdy enough for a useful dialogue with the results of relevance logic.

Is (ii) pertinent to an RA theorist? On the face of it, the answer is 'yes'. Suppose our RA theorist has settled on an account of relevance. There is then clear theoretical interest for her in developing a logical system where relevance is preserved across the proposed logical consequence relation. What is not so clear, however, is that the results of relevance logic provide a general enough framework for carrying out this job for arbitrary RA theorists, since, again, relevance logic is typically associated with a notion of relevance closely tied to preservation of *subject matter*. To the extent that interest in the RA approach fuels an interest in a diversity of accounts of relevance (see Sect. 7.3.4), it motivates a wider scope for relevance logic than mere attention to the interaction of consequence and subject matter.

In total: the basic concerns of relevance logic, at the very least, indicate an intriguing notion of relevance tied intimately to that of subject matter, an obvious matter of interest to the RA theorist. Beyond this, dialogue between relevance logic and the RA approach presents an intriguing, possibly fruitful but certainly subtle affair. Of course, our remarks are tentative: the relationship between relevance logic and RA theory deserves a more careful discussion.

7.2.9.2 Epistemic Relevance Between Evidence and Hypothesis

Let us now turn to a second tradition in philosophy in which the term "relevance" has received prominence. Here, the focus has been on when a piece of evidence is *relevant* to the evaluation of a hypothesis. There is therefore a parallel with the concerns of the relevant logician: while the relevant logician is concerned with when a conclusion genuinely follows from its premises, so one who investigates "epistemic relevance" in the present tradition is concerned with when evidence is genuinely a *reason* to accept (or reject) a hypothesis. The traditional starting point in this investigation has been a probabilistic account that states the following: evidence E is relevant to hypothesis H just in case the conditional probability of H given E is different to the (prior) probability of H. Discussion in the literature – initiated chiefly by Keynes and Carnap – has essentially developed as a series of refinements of this basic idea (Keynes 1921; Carnap 1950; Floridi 2008).

Analogously to the case of relevance logic, the discussion of "epistemic relevance" hinges on two basic concerns: (i*) what is the correct account of the relevance at issue? (ii*) How is this account to be integrated into a theory of evidential support? Once again, the extent to which the discussion of this sense of epistemic relevance relates to the concerns of the RA theorist depends on the extent to which answers to i* and ii* bear on these concerns.

On a first pass, we may make observations similar to those made with respect to the relationship between the RA approach and relevance logic. With respect to i*: the notion of relevance at work in the discussion of "epistemic relevance" is of interest to the RA theorist insofar as it represents, surely, *one* intriguing candidate for the notion of relevance the RA theorist believes can be identified as at work in the theory of knowledge (namely, a candidate that appeals to notions of *probability* and *independence* as crucial features). It remains to be seen, however, how far such a version of RA theory could be developed with plausibility. With respect to ii*: again, the RA theorist certainly ought to have interest in any general techniques for integrating an account of relevance into a theory of reasoning or evidential support (perhaps in aid of a *relevant alternatives theory of justification* that underlies the RA theory of knowledge). The apparent focus in the "epistemic relevance" literature on a quite specific notion of relevance does not inspire hope, however, that very *general* tools for such integration are to be found there. Once again, however, our remarks cannot be understood as anything other than preliminary, and clearly a deeper investigation is a worthwhile task.

7.2.9.3 Methodology of Science

Various strands in the literature on the epistemology and methodology of science have, it seems to me, a notable *prima facie* affinity to the ideas animating the RA theorist (not forgetting, of course, the motivation for the RA approach as a universal solution to under-determination problems, discussed in Sect. 7.2.5). We sketch a few such points.

Begin with the plausible idea that the *goal* of scientific inquiry is knowledge. If then we agree with the RA theorist that knowledge is always relative to a set of relevant alternatives, it seems that we should conclude that the methods of scientific inquiry – geared towards producing such knowledge – themselves operate against the backdrop of a set of relevant alternatives. If so, we should expect a notion of "relevant alternative" to play a role in both the *context of justification* and *context of discovery* of scientific hypotheses.

Indeed, ideas of roughly this ilk have received considerable attention in the literature. A linchpin of discussion in the philosophy of science literature of the last few decades has been Kuhn's proposal that major developments in the history of science amount to revolutionary upheavals brought about by a fundamental shift in underlying "paradigm" for normal science (Kuhn 1970). We need not here become engrossed in the substantive or scholarly issues connected to Kuhn's work. We need only note a potential for an RA approach to offer tools to understand and investigate such revolutionary shifts: for an RA theorist, a change in paradigm, it might be suggested, involves a major shift in the space of the *relevant hypotheses* that a normal scientist must seek to select between.

More specifically, let us consider the *context of justification*. In recent years, there has been a revival of the idea that scientific justification essentially amounts to a process of *eliminative induction* (cf. Earman (1992, Ch.7)): roughly, given a space of relevant hypotheses H_1 through H_n, a particular hypothesis H_i is supported by scientific inquiry just in case the available evidence rules out every competing hypothesis. One reason to initially find such an account of scientific methodology to be philosophically naive is to point out both the unwieldiness of the space of logically possible hypotheses, and the inability of our actual evidence to rule out any significant portion of this space (a version, of course, of *under-determination problem*). A successful RA theory, however, will presumably provide a notion of relevance that essentially defuses both problems. RA theory seems, therefore, a potential ally to the eliminative inductivist.

Turn now to the *context of discovery*. For the RA theorist, a natural way to understand the *discovery* of a new hypothesis is for that hypothesis – through whatever mechanism – to become *relevant* in the context of scientific inquiry. Whether this mechanism is rational or not will depend on the exact account of what relevance is, and how the space of relevant alternatives might change. In this connection, consider Hintikka's recent work in developing a logic of discovery that, roughly, posits scientific discovery as essentially amounting to posing a question to a source of information in nature (Hintikka 1999). On such an approach, scientific inquiry depends crucially on the *questions* that are (implicitly or explicitly) asked

by scientists. To the extent that *background questions* may therefore be understood as a source of *relevant alternatives* (see our discussion of *question-first RA theory* in Sect. 7.4 of this article), we see yet another venue for potential convergence between RA theory and important debates in the methodology literature.

7.3 Choice Points for the RA Theorist

In this section, we make a start at developing the RA approach with technical precision.

First, we present a 'minimal' RA theory. This theory is minimal insofar as it operates at a high level of abstraction yet, we claim, captures the core elements of the approach. In the process, we lay out a logical language that we work with throughout the rest of the paper, and a basic semantics for this language.

The minimal approach is too abstract to engage fully with philosophical debate. Likewise, its abstractness precludes it from encompassing the interesting formal features of more concrete RA theories. To this end, we discuss the potential for considering precise RA theories with more content. With this in mind, we next list a number 'choice points' for the RA theorist, by which one may divide the family of RA theories into a large number of species. To settle each choice point is to arrive at a 'concrete' RA theory.

It is worth emphasizing two points concerning the philosophical motivation for the formal work that follows. First, as is often the case with highly abstract frameworks, minimal RA theory holds limited theoretical interest in itself. Rather, it is a unifying skeleton upon which to hang the features of more concrete RA theories. Nevertheless, an important philosophical point is attached to our presentation of a minimal theory: at its most abstract, the RA approach is *very general*, a point for critics to keep in mind when aiming for blanket objections to the approach. Our second point is an acknowledgement that, from a philosophical point of view, the utility of a move to the formalities of a logical approach – with its accompanying idealization and austerity – is to be judged by its pay-off for the perspicuous study and presentation of philosophically relevant results. We hope to demonstrate modest but genuine results along these lines in Sect. 7.4.

7.3.1 An Epistemic Language

Let **At** be a set of atomic proposition letters. We work with the following logical language $\mathscr{L}$:

$$\varphi \quad ::= \quad p \mid \neg\varphi \mid \varphi \wedge \varphi \mid K\varphi \mid R\varphi \mid I\varphi \mid [\varphi]\varphi$$

where $p \in \mathbf{At}$.

The rest of the connectives are defined as usual. $K\varphi$ is intended to mean "the agent knows that φ". $R\varphi$ is intended to mean "φ is relevant". $I\varphi$ is intended to mean "the agent has the information that φ". The intended interpretation of $[\varphi]\psi$ is "after the set of relevant propositions is updated so as to be relative to φ, ψ is true". This last expression represents a *dynamification* of our logic (van Benthem 2011) (and we fully intend to draw on the techniques of this area – which includes the likes of well-studied dynamic epistemic logics such as *public announcement logic* – in our development).

We may then define a *two-place* relevance operator: $\mathtt{R}(\varphi, \psi) ::= [\varphi]R\psi$. The aim here is to capture the idea that ψ is relevant (perhaps only) relative to proposition φ.

7.3.2 Minimal RA Theory

In what follows, $\mathscr{P}(A)$ refers to the power-set of set A.

Definition 7.1 (Minimal RA model). A *minimal RA model* is a tuple

$$\langle W, \{\mathbf{R}_w\}_{w\in W}, \{E_w\}_{w\in W}, \{*_w\}_{w\in W}, V\rangle$$

where,

- W is a set of *points of evaluation*. The reader may think of these as "possible worlds", subsets of which are "unstructured propositions".
- $\mathbf{R}_w \in \mathscr{P}(\mathscr{P}(W))$ is a set of sets of worlds i.e. a set of propositions. This is the set of *relevant propositions* at world w.
- $E_w \in \mathscr{P}(W)$ is a set of worlds i.e. a proposition. This is the agent's *total evidence* or *total information* at world w.
- $*_w$ is an *update operation* accepting a sentence $\varphi \in \mathscr{L}$ and returning an updated model we denote by $\mathscr{M} *_w \varphi$. We stipulate that the only distinction between $\mathscr{M}$ and $\mathscr{M} *_w \varphi$ lies in the relevant propositions.
- V is a valuation assigning atoms to worlds.

Given minimal RA model $\mathscr{M}$ and world w, define the set $\mathbf{U}_w$ as follows:

$$\mathbf{U}_w = \{A \subseteq W \mid A \in \mathbf{R}_w \text{ and } A \cap E_w \neq \emptyset\}$$

Call $\mathbf{U}_w$ the set of *uneliminated propositions* at w: the set of propositions that are both relevant and compatible with the agent's evidence at w.

Two remarks are in order. Though we use the 'worlds' terminology to talk about our points of evaluation, there is no technical necessity attached to this interpretation. One may equally well talk about scenarios, centered worlds, or so forth. (Though, it should be remarked, the totality of the propositional valuations associated with each world would make one hesitate to think of them as mere 'situations'.)

Second, we deliberately leave it vague how exactly to interpret E_w. Is this the evidence that the agent has access to *in principle*, though she may not in fact *have*

all this evidence in her possession? Is this a conjunction of the individual pieces of evidence at the agent's disposal? If this is the agent's *information*, are we to understand E_w as being a *true* proposition i.e. $w \in E_w$? Indeed, settling these questions might indicate the unsuitability of representing E_w as a proposition, as opposed to, say, a set of (possibly incompatible) propositions. We say little to settle such questions in the present paper.

We now turn to semantics. As is typical in epistemic logic, it is worth bearing in mind that this semantics is best understood as describing an *idealized* agent. Idealized in what sense? For our purposes, we may understand our agent as follows: for our agent, relative to her information, there is no distinction between actual (explicit) and potential (implicit) knowledge. An ordinary human has much implicit knowledge relative to her information: knowledge that *could* be acquired by correct reasoning from the information and knowledge she already holds and yet, for whatever reason, she has not in fact acquired this knowledge. Our idealized agent has no such limitation.

Idealization raises the question: what is the relationship between our ideal agents and ordinary human beings? In particular, why think that a logical analysis of the one will shed light on the other? This is a subtle issue that deserves more discussion than we can give it here. One or two quick suggestions as to the relevance of ideal agents might prove useful to the reader, however. First, presumably, the study of any epistemic *limitation* of our ideal agents will have bearing on ordinary agents, since the former will also face that limitation. This observation seems particularly pertinent when it comes to under-determination problems, since, presumably, our ideal agents have no special advantage in terms of the empirical evidence at their disposal. Second, the move to idealization allows for elegant simplifications of certain issues. For instance, we will later meet certain proposed principles of epistemic closure that are difficult to state in full generality for ordinary agents, but simple to state in the case of ideal agents.

In what follows, read $[[\varphi]]_{\mathscr{M}}$ as:

$$[[\varphi]]_{\mathscr{M}} = \{w \in W \mid \mathscr{M}, w \vDash \varphi\}$$

Definition 7.2 (Minimal RA semantics). Given a minimal RA model $\mathscr{M}$, we define satisfaction at world w as follows:

- $\mathscr{M}, w \vDash p$ just in case $p \in V(w)$.
- $\mathscr{M}, w \vDash \neg\varphi$ just in case $\mathscr{M}, w \nvDash \varphi$.
- $\mathscr{M}, w \vDash (\varphi \wedge \psi)$ just in case $\mathscr{M}, w \vDash \varphi$ and $\mathscr{M}, w \vDash \psi$.
- $\mathscr{M}, w \vDash R\varphi$ just in case $[[\varphi]]_{\mathscr{M}} \in \mathbf{R}_w$.
- $\mathscr{M}, w \vDash I\varphi$ just in case $E_w \subseteq [[\varphi]]_{\mathscr{M}}$.
- $\mathscr{M}, w, \vDash K\varphi$ just in case $\{A \in \mathbf{U}_w \mid A \subseteq [[\neg\varphi]]_{\mathscr{M}}\} = \emptyset$.
- $\mathscr{M}, w \vDash [\varphi]\psi$ just in case $\mathscr{M} *_w \varphi, w \vDash \psi$.

Effectively, the clause for $I\varphi$ says that the agent has the information that φ just in case the agent's information entails φ. The clause for $K\varphi$ says the following: the agent knows φ just in case there is no proposition that entails $\neg\varphi$ that is uneliminated

i.e. both relevant and compatible with the evidence. The clause for $[\varphi]\psi$ simply says that such an expression is satisfied when ψ holds when the relevancy sets have been updated according to the $*_w$ operation with φ as input.[8]

A technical remark is in order. Our approach to the semantics of minimal RA theory clearly falls within the tradition of *neighbourhood semantics* for modal logic (Chellas 1980), where the truth clause for $\Box\varphi$ ("it is necessary that φ") amounts to: $\Box\varphi$ holds at world w just in case the set of worlds where φ holds is one of a set of "necessary propositions" associated with w. A technical elaboration of our proposed logic will therefore make use of the tools of neighbourhood semantics.

By now, the reader may well feel that we have left out something important in our RA account of knowledge. Presumably, knowledge is factive: if P is known, then P is true. What is more, knowledge implies belief – perhaps even *justified* belief. Yet none of these (some would say obvious) features are represented in our account of knowledge.

Incorporating these features into an RA account is a more subtle business than might first meet the eye. We illustrate the issues by remarking on the factivity of knowledge.

One option for the RA theorist is simply to add an additional component to the truth condition for knowledge: in addition, it must be the case that φ is *true* at w. This is of course a structural feature of countless proposed theories of knowledge.

Such a maneuver might strike those that wish to understand attaining a knowledge state as offering a *guarantee* of truth as somewhat ad hoc and unsatisfying. More satisfying, it might seem, would be an account of knowledge such that the conditions on knowledge attainment *non-trivially entail* truth. Indeed, the independence of a 'truth condition' from the other conditions in a theory of knowledge might be exactly what allows for the construction of the familiar Gettier cases that have dogged epistemology (Zagzebski 1994).

It might therefore be seen as an advantage of RA theory that it provides tools for ensuring that truth is entailed by knowledge without the stipulation of an independent truth condition. For all that would be required is that: any true proposition at w is relevant at w; and the proposition E_w lives up to the title of 'information' by in fact being a *true* proposition. Both proposals have some appeal: it might seem strange to deem the truth as irrelevant, and it might seem natural to insist that E_w constitutes the agent's *basic evidence* (say, her memories and immediate sensory experience (Lewis 1996)) and that such evidence must be compatible with the actual world. Of course, these quick remarks do not settle the matter.

The foci of our discussion in the rest of this paper means that we can generally safely put aside the issue of truth and justified belief, so we will for the moment simply fail to propose a way of incorporating these features. The reader is right to recognize the gap, however.

[8]The reader will note that we make no mention of a notion of 'context' anywhere in this semantics. We gloss over the role of context, as follows: context may be thought of as settling the valuation V and, potentially, the set of relevant alternatives $\mathbf{R}_w$. Thus, context may be thought of as settling the model in question. We do not explore this thought in any detail here.

7.3.3 Choice Points

We now turn to a series of choice points the RA theorist must settle in order to fill out the minimal approach. We summarize the points here, before elaborating in the coming subsections (though we first pause to better note my intentions in enumerating this list).

An RA theorist ought to ultimately answer the following questions:

- What is relevance?
- What sort of thing is an 'alternative'?
- What is it to 'rule out' an alternative?
- What are the *primitive* objects of relevance from which we derive relevance of an alternative?
- Does the (ir)relevance of a claim only make sense in *contrast* to another claim?
- Interaction principles: is relevance a necessary condition on knowing? Is irrelevance a sufficient condition for knowing the denial?

Some notes about this list of choice points. First, I do not intend to be understood as claiming that these choice points are entirely *independent* of one another. Indeed, we will see some instances (in Sect. 7.4) of how settling certain choices in a particular way constrains how other choices can be settled. Second, I do not claim that the manner in which an RA theorist can settle these choice points is entirely arational or arbitrary: there may well be good reasons for favouring one choice over another. Third, though I suspect this list is complete (in the sense that settling these issues produces a concrete RA theory), I will refrain from defending this point here.

7.3.4 Relevance

Perhaps the key philosophical matter that the RA theorist needs to settle is the question as to what *relevance* comes to.

This is no simple matter: the literature on RA theory displays a bewildering diversity of suggestive comments, but is light on detailed theories of relevance. Cohen (1999, p.61) suggests that relevance is a matter of the psychology of the agents in conversation, "determined by some complicated function of speaker intentions, listener expectations, presuppositions of the conversation, salience relations etc.". Heller (1999) suggests that relevance is a matter of similarity to the actual world, where the similarity relation is itself settled partially by psychological facts – intentions, salience and so forth – of the speakers in context. Lewis (1996) suggests a complex array of factors that determine relevance, ranging from salience to the speaker to practical stakes. In contrast, Dretske (1981) is somewhat non-committal, but indicates some commitment to the idea that relevance is a purely *objective matter*, independent of the agent's state of mind.

7.3.5 Alternatives

What is it for a 'circumstance' to be an alternative to a proposition? What sort of thing is an alternative?

Again, we deliberately use the term 'circumstance' here in a neutral manner. In our minimal model, we have effectively treated alternatives as *unstructured propositions* i.e. as a set of possible worlds. Though, again, we emphasize that in the minimal model, one need not read too much into this choice of terminology: the "worlds" are simply points of evaluation. At any rate, to treat alternatives as propositions is a typical move in the literature, as is the following definition: A is an alternative to P just in case A entails the negation of P.

But this is not the only option. An alternative could be modeled as a *situation* or *set of situations* (Barwise and Perry 1983), where formally a situation is akin to a possible world, only that a valuation on situations can be partial. Along similar lines, an alternative could be understood as a *structured proposition*. Other options include: as a centered proposition; as an interpreted sentence; and perhaps others. The choice here might well require more than simply adding fine structure to minimal RA models, but call for perhaps more radical variations on the minimal models and their semantics. For the purposes of this essay, we continue to treat alternatives as unstructured propositions – but this choice is more for convenience than principle.

One further option as to what sort of thing we might take an 'alternative' to be is worth focusing on momentarily: instead of thinking of an alternative as a proposition (of some sort), we could instead think of alternatives as *possible worlds*. Namely, world w is an alternative to proposition P just in case $\neg P$ holds at w. Indeed, there is, in my opinion, a great deal of *ambiguity*, between thinking of alternatives as propositions or as worlds, in some of the key philosophical texts in the RA literature (Dretske 1981; Lewis 1996).

We can, however, capture the stipulation that alternatives are worlds within our current framework with an appropriate restriction: that only *singleton sets* can occur as relevant propositions. With such a stipulation, the truth clause for $K\varphi$ essentially amounts to: $K\varphi$ holds at world w just in case every relevant world u in which $\neg\varphi$ holds is not a member of the agent's evidence set i.e. is incompatible with the agent's evidence.

7.3.6 Ruling Out

What is it for an agent's evidence to *rule out* an alternative?

In our minimal models, we captured a notion of a proposition P being *incompatible* with the agent's evidence E (E itself understood as proposition): namely, that the intersection of P and E is empty i.e. at no world is the evidence true and yet P is false. This notion of incompatibility serves as one natural and basic attempt at

capturing the idea of ruling out an alternative, roughly capturing Dretske's idea of having a reason that is *conclusive* with respect to P.[9]

This notion of incompatibility may be substantiated in different ways, however. For one thing, the matter as to how exactly to interpret the set E_w – and whether a set of worlds is at all the best modeling device for this purpose – is a delicate issue (briefly touched upon in our discussion of the minimal semantics in Sect. 7.3.2).

It might be thought, however, that mere incompatibility with the agent's evidence, while clearly *sufficient* for ruling out a proposition, could stand to be supplemented with a richer construal of 'ruling out'. I will emphasize what seem to me two important ways of developing this suggestion. One route is to connect the notion of 'ruling out' with the agent's rational belief: an alternative A is ruled out for the agent just in case it is rational for the agent to find A implausible. Call this the *soft approach* to ruling out. Since the formal treatment of rational belief is itself quite well developed (see, for instance, van Benthem 2011, Ch.7), tools for the integration of this approach into our formal model are available.

A second approach is to give 'ruling out' a stronger reading: for A to be ruled out for the agent is for the agent to *know* that A is false. Call this the *hard approach* to ruling out. This account has an interesting consequence: if A being an alternative to P means that P entails $\neg A$, then, on the current view, the RA slogan is best understood as commitment to the idea that an agent can know P without knowing all of its entailments. This, then, amounts to a commitment to the denial of knowledge under entailment, even for our cognitively ideal agents, a constraint that can be captured in logical terms.

Thus, the choice between the soft approach and hard approach is intimately tied to the debate concerning the status of epistemic closure principles.

7.3.7 The Primitive Objects of Relevance

We now consider some less obvious choice points for the RA theorist.

We have been speaking about the relevance of *alternatives* (understood here as propositions). But certain developments in the literature indicate that it is worth considering the relevance of alternatives as *derived* from the relevance of some more fundamental kind of object to which relevance applies.

Heller,[10] for instance, suggests that it is *possible worlds* that are the primitive objects of relevance (Heller 1989, 1999): in a context, some worlds are *similar enough* to the actual world to be considered relevant. Call this the *worlds-first*

[9] It also goes some way towards capturing Lewis' notion of ruling out (Lewis 1996): for him, A is ruled out just in case it holds at no possible world in which the agent has the same memories and sensory experience.

[10] We may want to place Dretske and Nozick in this camp too.

approach. How then do we recover the relevance of propositions? As follows: A is relevant just in case A holds at some relevant world.

Switching to possible worlds as the objects of relevance is a simple formal matter: we could, for instance, no longer treat $\mathbf{R}_w$ as a set of propositions, but instead as a set of worlds. We may then alter the satisfaction clause for knowledge as follows:

$$\mathscr{M}, w \vDash K\varphi \text{ just in case } E_w \cap \mathbf{R}_w \subseteq [[\varphi]]_{\mathscr{M}}$$

On the other hand, Jonathan Schaffer has, in a series of recent papers (Schaffer 2004, 2005a, 2007a,b), proposed that knowledge claims can only be evaluated relative to a background *question*. Call this the *question-first approach*. To know something, according to this idea, is to know it *rather than* other possible answers to the question, while non-answers and presuppositions to the question are simply ignored. The idea is that in answer to the question "is there a zebra in the cage or nothing at all?" one may know that there is a zebra, but in answer to the question "is there a zebra in the enclosure or a painted mule?" one may not know that there is a zebra. From an RA perspective, this suggests a notion of relevance for propositions: A is relevant relative to (relevant) question Q just in case A is an *answer* to Q. Fortunately, there is an ongoing tradition in the semantics literature from which to draw for treating questions formally (Hamblin 1958, 1973; Belnap and Steel 1976; Ciardelli et al. 2013). For our immediate purposes, we may understand a question Q, in the formal sense, as a set of disjoint propositions, representing the set of (least specific) answers to that question. An answer to the question is then any subset of a member of Q. A *partial* answer is any *union* of subsets of Q. A presupposition to the question is any proposition that contains every member of Q.

According to another approach, Stephen Yablo has, again in recent work (Yablo 2014), proposed that knowledge claims can only be evaluated relative to a background *subject matter* or *topic (of conversation)*. Call this the *topic-first* approach. On this view, one can know that the enclosure contains a zebra so long as the subject of painted mules is suppressed. From an RA point of view, this suggests a notion of relevance for propositions as follows: A is relevant relative to (relevant) topic T just in case A concerns (only) that relevant subject matter. Again, fortunately, there are formal tools for integrating subject matters into a formal setting (Lewis 1988): a subject matter, according to Lewis, can be understood as a *partition* on the space of possible worlds, with two worlds sharing a cell just in case they are exactly the same when it comes to any state of affairs concerning that subject matter.

7.3.8 Contrast

Is the relevance of a claim A a notion that only makes sense relative to another claim, to which A is to be *contrasted*? Let us say that a theory that answers this question in the affirmative takes the *contrast approach*.

Dretske's theory (Dretske 1970, 1981) subscribes to the contrast approach, so will serve as a useful illustration. For Dretske, it does not make sense to describe proposition A as relevant or irrelevant independent of a proposition to which it is to be contrasted. Rather, A can only be understood as relevant or not when understood as an *alternative* to a particular proposition P (in which case, we can settle whether A might be true were P to be false, following Dretske's notion of relevance).

On the other hand, Lewis' theory (Lewis 1996) does not subscribe to the contrast approach. For Lewis, once the context is fixed, a proposition A is uniformly relevant (or not), no matter which proposition P is being evaluated for knowledge.

Subscribing to the contrast approach, perhaps surprisingly, has far-reaching consequences for an RA theory: Holliday (2014, 2015) shows that subscription to the contrast approach is the source of the closure failures exhibited by Dretske's theory, while resisting the contrast approach is essentially exactly what allows Lewis to preserve closure in his own theory.

For the purposes of formalization we can capture taking the contrast approach as follows: we may define $\mathtt{K}\varphi$ ('proper knowledge') as follows: $\mathtt{K}\varphi ::= [\varphi]K\varphi$ i.e. proper knowledge of φ is understood as 'knowledge' of φ in the wake of an update that relativizes relevance to φ. That is, we incorporate the contrast approach by stipulating that evaluation of a knowledge claim involves an *update* of the relevancy set (cf. Holliday 2012).

7.3.9 *Interaction Principles Between Relevance and Knowledge*

In terms of logical principles, what relationship should exist between the relevance of a proposition and knowledge of that proposition? Should there be no such logical relationship? Should the relevance of A act as a necessary condition on knowledge of A (that is, should only *relevant* propositions count as candidates for knowledge)? Should the irrelevance of A be sufficient for $\neg A$ to be known, or for A to be *not* known?

In terms of integration into our framework, stipulating that relevance be a necessary condition on knowledge is at least a simple matter: we simply add the condition $\mathscr{M}, w \vDash R\varphi$ to the clause for $K\varphi$.

7.4 Survey and Classification of Representative RA Theories

We now exhibit a sample of RA theories, making use of the choice points from Sect. 7.3 to build some interesting (still relatively abstract) theories that relate to recent and important discussions in the epistemology literature. The first three choice points are the most obvious choice points, and also the hardest to get a formal

grip on. For interest and convenience, we essentially focus on the last three: the choice of primitive objects of relevance; the choice between adopting and rejecting the contrast approach; and the choice as to whether to treat relevance as a necessary condition on knowledge.

One goal of this section is to simply exhibit the diversity of RA theories. Another is to illustrate the difference that settling certain choice points can make, and the implications for other choice points. Another is to demonstrate the formalization of RA theory in action, and demonstrate how formalization can substantively sharpen and otherwise contribute to the philosophical debate.

To achieve this last end, it will be of interest to consider three principles that we can express in our language, and that have bearing on the philosophical evaluation of an RA theory:

- $K\varphi \rightarrow K\psi$ whenever $\mathscr{M} \vDash \varphi \rightarrow \psi$ **(Closure under entailment)**
- $K\varphi \wedge K(\varphi \rightarrow \psi) \rightarrow K\psi$ **(Closure under known implication)**
- $K(\varphi \wedge \psi) \rightarrow K\varphi \wedge K\psi$ **(Conjunctive distribution)**

By $\mathscr{M} \vDash \varphi$ we mean the standard thing: that φ is true at every world in $\mathscr{M}$.

What is notable about the first two principles is that their validity is *philosophically controversial* (and so where a theory lands on the validity of these principles has philosophical significance) (Luper 2001). Recall Dretske's famous example and diagnosis: one may know that the animal in the enclosure is a zebra, without knowing that it is not a painted mule, even though it being a zebra entails that it is not a painted mule. Undoubtedly, Dretske has zeroed in on an important feature of our intuitive judgements. Yet, on the other hand, there are reasons to *resist* dropping closure under known implication: the validity of this principle, it might be said, represents the fact that deductive reasoning from known claims is always a source of knowledge, at least given the idealizations we are working with (cf. Kripke (2011)).[11]

What is notable about the third principle above is that it is *not* controversial (Kripke 2011; Yablo 2014). That a theory invalidates conjunctive distribution may therefore be understood as an unequivocal *strike* against that theory.

7.4.1 Examples of the Question-First Approach

Let us briefly explore some variations on the question-first approach. Schaffer's work, again, is not explicitly located within the RA tradition,[12] but there is nothing

[11]For recall that we self-consciously model the knowledge of an ideal agent that is always able to "put two and two together" and can therefore maximally extend her knowledge by way of reasoning. To deny of such an agent that closure under known implication is valid is to deny that we ordinary agents are always in principle able to extend our knowledge using self-conscious deductive reasoning by way of known implications.

[12]Though perusal of, for instance, Schaffer (2005a,b) quickly reveals the close ties between Schaffer's views and the RA approach.

stopping an RA theorist from viewing it – or at least certain borrowed aspects of it – in this light. I make no claim in what follows to be representing the details of Schaffer's work entirely accurately (for that, I direct the reader to Schaffer 2004, 2005a, 2007a,b). In the name of convenience, our aim is to operate according to its spirit, not its letter.

Again, we formally understand a question Q as a disjoint (but not necessarily exhaustive) set of propositions, representing the least specific distinct answers to that question. We integrate this into our RA model as follows:

Definition 7.3 (Interrogative RA model). An *interrogative RA model* is a tuple

$$\langle W, \{\mathbf{Q}_w\}_{w \in W}, \{E_w\}_{w \in W}, V \rangle$$

where,

- Every element is as in a minimal RA model, except:
- $\mathbf{Q}_w$ is a set of disjoint propositions.

We first present an RA theory based on these models that rejects the contrast approach (we shall say it is *contrast free*) and does not treat relevance as a necessary condition on knowledge. To do so, we do not need to alter the semantic clauses for that theory: we simply need to define the set $\mathbf{R}_w$ of relevant propositions. First: define $\mathbf{Q}_w^+$ – the set of all answers to $\mathbf{Q}_w$ – as follows

$$\mathbf{Q}_w^+ = \{A \subseteq W \mid A \subseteq A' \in \mathbf{Q}_w\}$$

Then:

$$\mathbf{R}_w ::= \{P \subseteq W \mid P = \bigcup \mathbf{A} \text{ for some } \mathbf{A} \subseteq \mathbf{Q}_w^+\}$$

That is, a proposition is relevant just in case it is an answer (or a partial answer) to the question $\mathbf{Q}_w$.

Effectively, our semantics for $K\varphi$, as detailed in Sect. 7.3.2, then turns out as follows: $K\varphi$ holds just in case φ holds throughout the partial answers to question $\mathbf{Q}_w$ that are not incompatible with the agent's evidence.

The following result is straightforward to prove. We leave the proof as an exercise for the reader.

Proposition 7.1. *The above RA theory*

- *validates closure under entailment;*
- *validates closure under known implication;*
- *(and therefore) validates conjunctive distribution.*

By virtue of the validity of closure under entailment, we may therefore note that the current RA theory cannot be one such that ruling out is understood as knowing the negation. So our RA theory is constrained to follow the *soft approach* to ruling out.

Let us now try a variation on the question-first approach: we leave consideration of the contrast approach to another time, but add to our theory that relevance is a necessary condition on knowledge. In the current context this says: one can only know P in response to question Q if P is in fact a (partial) answer to Q. Presuppositions to Q and other non-answers cannot be known – not as a matter of insufficient evidence perhaps, but since these do not *qualify as candidates* for knowledge in the context of the question at issue.

We therefore alter our semantics as follows:

$$\mathscr{M}, w \vDash K\varphi \text{ just in case } \mathscr{M}, w \vDash R\varphi \text{ and } \{A \in \mathbf{U}_w \mid A \subseteq [[\neg\varphi]]_{\mathscr{M}}\} = \emptyset$$

Proposition 7.2. *Our latest RA theory*

- *invalidates closure under entailment;*
- *invalidates closure under known implication;*
- *invalidates conjunctive distribution.*

One persuaded of the wisdom of rejecting closure will find our altered theory more amenable. But at a cost: conjunctive distribution is lost.

7.4.2 Examples of the Topic-First Approach

Let us now consider some versions of the topic-first approach. Again, we take inspiration from the work of Yablo, without here attempting to capture the intricate details of his full position (cf. Yablo 2014).[13]

Following Lewis (1988), we understand a topic T as a partition on the set of worlds W. The general idea is this: a topic amounts to a collection of *ways a world could be with respect to that subject matter*, providing an equivalence relation between worlds (two worlds are equivalent just in case they are indistinguishable with respect to how things are with respect to the topic in question).[14] For instance, if the topic is the seventeenth century (Lewis' example), then two worlds reside in the same cell of the partition associated with this subject matter just in case affairs with respect to the seventeenth century are identical in those two worlds.

In the setting of a propositional logic, it is convenient and somewhat natural to instead define a topic T as a set of *interpreted atomic proposition letters* (cf. the semantics of *relatedness logic* (Epstein 1994; Burgess 2009)). This then serves to *define* a partition on the space W: two worlds w and w' are equivalent just in case they are agree on the truth value of each atom in T. Call this partition π_T.

[13] We mention in footnotes some divergences from important details of Yablo's theory as we proceed.

[14] Yablo in fact embraces a more general conception of a topic: for him, a topic can be associated with a *reflexive, symmetric* relation on the space of worlds, as opposed to an equivalence relation.

It is worth remarking on the nature of the partition that a subject matter invokes. We may think of this partition (if non-trivial) as imposing a coarser *resolution* on the space of possible worlds, whereby two possible worlds are treated as indistinguishable unless they differ with respect to the state of the subject matter in question. A subject matter, then, may be understood as controlling the distinctions that are recognized in the space of possibilities: distinctions involving the subject matter are visible, while those that 'cut across' the subject matter are invisible.

Definition 7.4 (Topical RA model). A *topical RA model* is a tuple

$$\langle W, \{\mathbf{T}_w\}_{w\in W}, \{E_w\}_{w\in W}, V\rangle$$

where,

• Every element is as in a minimal RA model, except:
• $\mathbf{T}_w$ is a set of atoms, for each $w \in W$.

Let us again begin with an RA theory that is contrast free and does not impose relevance as a necessary condition on knowledge. We may then define the set of relevant propositions as follows:

$$\mathbf{R}_w ::= \{P \subseteq W \mid P = \bigcup \mathbf{C} \text{ for some } \mathbf{C} \subseteq \pi_{\mathbf{T}_w}\}$$

That is: a relevant proposition is identical to a union of cells in the partition determined by the given topic. Other propositions are irrelevant, as they involve distinctions that are 'invisible' to the given subject matter.

We can think of the information the agent's evidence E delivers *with respect to the current subject matter* T as amounting to the smallest union of cells in π_T that contains E. Label this union E^T. Now: given topical RA model $\mathscr{M}$ and world w, define the set $\mathbf{U}_w$ as usual:

$$\mathbf{U}_w = \{A \subseteq W \mid A \in \mathbf{R}_w \text{ and } A \cap E_w \neq \emptyset\}$$

Our semantics are also as before. The net result: $K\varphi$ holds just in case φ holds throughout every cell of $\pi_{\mathbf{T}_w}$ contained in the topic-relevant information $E_w^{\mathbf{T}_w}$. It may once again be checked that the following hold:

Proposition 7.3. *The above RA theory*

- *validates closure under entailment;*
- *validates closure under known implication;*
- *(and therefore) validates conjunctive distribution.*

It might strike the reader that our last RA theory does not have much appeal, potentially pleasant formal features aside: since $E_w \subseteq E_w^{\mathbf{T}_w}$, there is little epistemic advantage for the ideal agent that adopts a restricted subject matter over consideration of the whole of logical space (for – perhaps intuitively! – refining the relevant topic on this theory tends to *improve* the informational situation of the

agent, as this allows her to discern distinctions that were previously ignored). So let us consider one last topic-first theory, one that moves a little closer to Yablo's own presentation: namely, a topic-first approach that both embraces the contrast approach and offers a more nuanced view as to when a proposition is incompatible with the agent's information. That is, in the following theory, relevant subject matter – and so the relevance of propositions – is fixed relative to whatever proposition is being evaluated for knowledge. To accomplish this, we need to provide a fleshed out semantics for $[\varphi]\psi$ expressions.

The idea will be as follows: every sentence φ in the language embodies a natural subject matter: the set T^{φ} of atoms that occur in φ.[15] Our update operator $[\varphi]$ will simply update current model $\mathscr{M}$ so that $\mathbf{T}_w$ is replaced with T^{φ}, giving model $\mathscr{M} *_w \varphi$. The semantic clause is as follows:

$$\mathscr{M}, w \vDash [\varphi]\psi \text{ just in case } \mathscr{M} *_w \varphi, w \vDash \psi$$

We now consider a more nuanced account as to when a proposition is incompatible with the agent's information.

Definition 7.5 (Ordered topical RA model). An *ordered topical RA model* is a tuple

$$\langle W, \{\leq_w\}_{w \in W}, \{\mathbf{T}_w\}_{w \in W}, \{E_w\}_{w \in W}, V\rangle$$

where,

- Every element is as in a topical RA model, except:
- $\leq_w$ is a total order on W, with w a minimal element in the ordering.

Think of $\leq_w$ as a measure of *distance from world w* on the worlds W. We will make use of this ordering to capture a notion as to when evidence E is a conclusive reason for rejecting P: namely, this is the case exactly when in the *nearest worlds* to actuality in which P is true, E is false (cf. Dretske 1971). Then, we deploy the following idea: P is incompatible with E just in case E is a conclusive reason for rejecting P.[16]

Given proposition A and world $u \in A \subseteq W$, we say that u is $\leq_w$*-minimal with respect to* A just in case there is no world in A closer to w than u, according to $\leq_w$. With this in mind, given ordered topical model $\mathscr{M}$ and world w, define the set $\mathbf{U}_w$ as follows:

[15] Here we see another divergence from Yablo. For Yablo, the subject matter associated with φ is the set of *ways that* φ *could be true and the ways it could be false*. More precisely: it is the set of (what Yablo calls) the *minimal partial models* that succeed in either making φ true or φ false. The reader interested in a proper explication of these notions is directed to Yablo (2014).

[16] We depart from Yablo here as follows: for Yablo, elimination of alternatives is inspired by the "tracking" account of Nozick: A is eliminated just in case the agent believes $\neg A$ and were A to be the case, the agent would *not* believe $\neg A$. Despite this change in perspective, the technical details for Yablo's account and our own are similar in many respects.

$$\mathbf{U}_w = \{A \subseteq W \mid A \in \mathbf{R}_w \text{ and } \exists u \in \text{A s.t. u is } \leq_w \text{-minimal w.r.t A and } u \in E_w\}$$

That is: a relevant alternative A to P is now understood to be eliminated by evidence E just in case E is a conclusive reason for rejecting A. A is uneliminated just in case A is relevant and E holds at some nearest A-world to actuality.

Otherwise, our semantic clauses are unchanged. However, we now define 'proper knowledge' of φ as $\mathtt{K}\varphi ::= [\varphi]K\varphi$ and relative relevance as $\mathtt{R}(\varphi, \psi) ::= [\varphi]R\psi$.

The net effect: $\mathtt{K}\varphi$ holds at w just in case for every cell C in $\pi_{T_w^\varphi}$ throughout which φ is false, E is a conclusive reason to reject C.

Though the complexity of our models is piling up, the following is still relatively easy to check (see Hawke for a comprehensive discussion):

Proposition 7.4. *Our latest RA theory (where we now understand the following principles in terms of* $\mathtt{K}$ *instead of* K*):*

- *invalidates closure under entailment;*
- *invalidates closure under known implication;*
- *validates conjunctive distribution.*

We remark that a key counter-example to closure on the current account is provided by principles of the form: $\mathtt{K}(p) \wedge \mathtt{K}(p \rightarrow p \vee q) \rightarrow \mathtt{K}(p \vee q)$.

This result in fact makes good on one of Yablo's leading motivations for considering knowledge relative to subject matter: the preservation of conjunctive distribution (as inference from a conjunction to a conjunct introduces no new subject matter) while discarding closure (since, according to Yablo, *disjunction introduction* can introduce new subject matter, and so the conclusion should be more elusive than the premises).

7.5 Conclusion

That concludes our whirlwind tour of the landscape of RA theories. We have accomplished the following: we have seen a number of informal philosophical motivations for embracing the RA approach, ranging from appeal to ordinary linguistic data to the drawing of lessons from famous philosophical examples; we have discussed a minimal framework for formalizing RA theory, and have considered at length various choice points that an RA theorist must decide upon in the construction of her theory; finally, we exhibited four RA theories, by way of setting some of the parameters from the previous section. At the same time, we drew important recent discussions of question and topic-relative knowledge into the RA fold, and demonstrated how the precision of logical techniques can be brought to bear on substantive issues of philosophical evaluation.

Acknowledgements I wish to thank Johan van Benthem and Krista Lawlor for much useful discussion during the development of material in this paper. Thanks for Chenwei Shi and an anonymous referee for many helpful comments based on a close reading of an earlier draft of this paper.

References

Anderson, A.R., Belnap, N.D.: Entailment: The Logic of Relevance and Necessity, vol. I. Princeton University Press, Princeton (1975)

Austin, J.L.: Other minds. Aristot. Soc. Suppl. **20**, 122–197 (1946)

Barwise, J., Perry, J.: Situations and Attitudes. MIT, Cambridge (1983)

Belnap, N.D., Steel, T.B.: The Logic of Questions and Answers. Yale University Press, New Haven (1976)

Burgess, J.P.: Philosophical Logic. Princeton University Press, Princeton (2009)

Carnap, R.: Logical Foundations of Probability. University of Chicago Press, Chicago (1950)

Chellas, B.F.: Modal Logic: An Introduction. Cambridge University Press, Cambridge/New York (1980)

Ciardelli, I., Groenendijk, J., Roelofsen, F.: On the semantics and logic of declaratives and interrogatives. Synthese **192**, 1689–1728 (2013)

Cohen, S.: How to be a fallibilist. Philos. Perspect. **2**, 91–123 (1988)

Cohen, S.: Contextualism, skepticism and reasons. Noûs **33**, 57–89 (1999). Supplement: Philosophical Perspectives, 13, Epistemology

DeRose, K.: Solving the skeptical problem. Philos. Rev. **104**(1), 1–52 (1995)

Dretske, F.I.: Epistemic operators. J. Philos. **67**(24), 1007–1023 (1970)

Dretske, F.I.: Conclusive reasons. Aust. J. Philos. **49**(1), 1–22 (1971)

Dretske, F.I.: The pragmatic dimension of knowledge. Philos. Stud. **40**(3), 363–378 (1981)

Earman, J.: Bayes or Bust?: A Critical Examination of Bayesian Confirmation Theory. MIT, Cambridge (1992)

Epstein, R.L.: The Semantic Foundations of Logic. Oxford University Press, New York (1994)

Floridi, L.: Understanding epistemic relevance. Erkenntnis **69**, 69–92 (2008)

Goldman, A.: Discrimination and perceptual knowledge. J. Philos. **73**, 771–791 (1976)

Hamblin, C.L.: Questions. Aust. J. Philos. **36**(3), 159–168 (1958)

Hamblin, C.L.: Questions in montague english. Found. Lang. **10**(1), 41–53 (1973)

Hawke, P.: Questions, topics and restricted closure, Philos. Stud. (forthcoming)

Hawke, P.: The problem of epistemic relevance. Ph.D dissertation, Stanford University (2016)

Heller, M.: Relevant alternatives. Philos. Stud. **55**(1), 23–40 (1989)

Heller, M.: Relevant alternatives and closure. Aust. J. Philos. **77**(2), 196–208 (1999)

Hintikka, J.: Inquiry as Inquiry: A Logic of Scientific Discovery. Kluwer Academic Publishers, Dordrecht (1999)

Holliday, W.H.: Epistemic logic, relevant alternatives and the dynamics of context. In: Lassiter, D., Slavkovic, M. (eds.) New Directions in Logic, Language and Computation. Lecture Notes in Computer Science, vol. 7415, pp. 109–129. Springer, Berlin/Heidelberg (2012)

Holliday, W.H.: Epistemic closure and epistemic logic I: relevant alternatives and subjunctivism. J. Philos. Log. **44**, 1–62 (2014)

Holliday, W.H.: Fallibilism and multiple paths to knowledge. In: Gendler, T., Hawthorne, J. (eds.) Oxford Studies in Epistemology, vol. 5, pp. 97–144. Oxford University Press, Oxford (2015)

Keynes, J.M.: A Treatise on Probability. Macmillan, London (1921)

Kripke, S.: Nozick on knowledge. In: Philosophical Troubles: Collected Papers, vol. 1. Oxford University Press, New York (2011)

Kuhn, T.S.: The Structure of Scientific Revolutions, 2nd edn. (1st edn. published 1962). The University of Chicago Press, Chicago (1970)

Lawlor, K.: Assurance: An Austinian View of Knowledge and Knowledge Claims. Oxford University Press, Oxford (2013)

Lewis, D.: Relevant implication. Theoria **54**(3), 161–174 (1988)

Lewis, D.: Elusive knowledge. Aust. J. Philos. **74**(4), 549–567 (1996)

Luper, S.: The epistemic predicament: knowledge, Nozickian track, and skepticism. Aust. J. Philos. **62**, 26–50 (1984)

Luper, S.: The Epistemic Closure Principle (2001). Substantive revision Wednesday 4 Aug 2010. http://plato.stanford.edu/entries/closure-epistemic/
MacFarlane, J.: The assessment sensitivity of knowledge attributions. In: Gendler, T., Hawthorne, J. (eds.) Oxford Studies in Epistemology, vol. 1, pp. 197–234. Oxford University Press, Oxford (2005)
Mares, E.: Relevance Logic (1998). http://plato.stanford.edu/entries/logic-relevance/. Substantive revision Mon 26 Mar 2012
Nozick, R.: Philosophical Explanations. Harvard University Press, Cambridge (1981)
Pritchard, D.: Epistemological Disjunctivism. Oxford University Press, Oxford (2012)
Schaffer, J.: From contextualism to contrastivism. Philos. Stud. **119**(1), 73–103 (2004)
Schaffer, J.: Contrastive knowledge. In: Gendler, T., Hawthorne, J. (eds.) Oxford Studies in Epistemology, vol. 1, pp. 235–271. Oxford University Press, Oxford (2005a)
Schaffer, J.: What shifts? Thresholds, standards, or alternatives?. In: Preyer, G., Peter, G. (eds.) Contextualism in Philosophy, pp. 115–130. Oxford University Press, Oxford (2005b)
Schaffer, J.: Closure, contrast, and answer. Philos. Stud. **133**(2), 233–255 (2007a)
Schaffer, J.: Knowing the answer. Philos. Phenomenol. Res. **75**(2), 383–403 (2007b)
Stanford, P.K.: Underdetermination of scientific theory (2009). http://plato.stanford.edu/entries/scientific-underdetermination/. Substantive revision Mon 16 Sep 2013
Stanley, J.: Knowledge and Practical Interests. Oxford University Press, New York (2005)
Unger, P.: Ignorance: A Case for Skepticism. Oxford University Press, New York (1975)
van Benthem, J.F.A.K.: Logical Dynamics of Information and Interaction. Cambridge University Press, Cambridge/New York (2011)
Vogel, J.: Are there counterexamples to the closure principle? In: Roth, M., Ross, G. (eds.) Doubting, pp. 13–27. Kluwer Academic, Dordrecht (1990)
Vogel, J.: The new relevant alternatives theory. Philos. Perspect. **13**, 155–180 (1999)
Yablo, S.: Aboutness. Princeton University Press, Princeton (2014)
Zagzebski, L.: The inescapability of Gettier problems. Philos. Q. **44**(174), 65–73 (1994)

Chapter 8
Knowledge Based on Reliable Evidence

Chenwei Shi

Abstract In this paper we propose to model each piece of evidence as a set of hypotheses that the evidence supports. By formalizing this idea, we can reason about knowledge based on the notion of "reliable belief". Our new understanding of knowledge highlights "reliability" of information that the agent gets from evidence. This is very different from the perspective of safe belief and its focus on "robustness". By a systematic comparison between these two kinds of beliefs, we argue that it is the reliability, not the robustness, that qualifies belief as knowledge. Finally, we explore the agent's knowledge update, particularly triggered by evidence dynamics, and present a complete dynamic logic.

Keywords Evidence as a set of hypotheses • Knowledge as reliable belief • Knowledge update and evidence dynamics

8.1 Introduction

Evidence has been the central concern throughout the history of epistemology. It plays an important role in our understanding of the notion of knowledge (Conee and Feldman 2004; Dougherty 2011). Recent study in epistemic logic brought this concept, as well as its relationship with belief and knowledge to logicians' attention. van Benthem and Pacuit (2011a) models an agent's body of evidence as sets of possible worlds, and each evidence set represents information that the agent has evidence for. Differently, Baltag et al. (2014) follows the tradition of justification logic (Artemov 2008), in which evidence terms are used to denote different pieces of evidence themselves, atomic or compound. In the same logic, the *evidential support relation* between evidence and hypotheses is expressed as $t \gg \varphi$, which reads as "evidence t is admissible for φ". And whether one piece of evidence is admissible for a proposition is presumed directly at the meta-level syntactically.

C. Shi (✉)
Tsinghua University – University of Amsterdam Joint Research Centre for Logic, Beijing, China

J. Redmond et al. (eds.), *Epistemology, Knowledge and the Impact of Interaction*, Logic, Epistemology, and the Unity of Science 38, DOI 10.1007/978-3-319-26506-3_8

In this paper, we follow the approach of van Benthem and Pacuit (2011a), modelling evidence in some logical space, semantically. In our view, though evidence sets as proposed in van Benthem and Pacuit (2011a) are enough for defining the notion of belief, this is not the case for knowledge. The following example illustrates our motivation:

Example 8.1.1. According to some authorized medical record, there is only one kind of disease A that can cause symptom S; and there are two kinds of diseases B and C that can cause a different symptom T. Now there are two patients, one with symptom S that has disease A, and the other with symptom T that has disease B. The doctor diagnoses the first patient as having disease A by symptom S and the second patient as having disease B by symptom T.

We tend to take the doctor's first diagnosis, not his second diagnosis, as knowledge, although both of them seem to be right. The reason is obvious, symptom S as evidence can rule out all the alternatives to disease A, whereas symptom T cannot. It is hard to judge whether the doctor's diagnosis can be counted as knowledge only by examination of the diagnosis alone. It also depends on the relationship between the symptoms and the diseases. It is this relation, the evidential support relation between a piece of evidence and hypotheses that cannot be captured by the evidence set. Therefore, a finer structure is called for to represent this aspect of evidence, so that the hypotheses supported by the agent's evidence can play an essential role. Toward this end, the current paper will propose a new evidence structure. We then discuss under what conditions the agent's "diagnoses" can be counted as knowledge.[1] In addition, we compare our new notion of evidence-based knowledge with the notion of safe belief (Baltag and Smets 2008) and show that our notion of knowledge takes a totally different direction pointed out by Stalnaker (2006), highlighting "the causal sources of beliefs" rather than robustness of belief.

More importantly, we will also take a dynamic perspective and look at knowledge update. In this context, our focus will be the knowledge update caused by the agent's evidence change. Typical questions to answer in the above example are the following: How would the doctor's knowledge change, if he were informed that disease C could also cause symptom T besides B? Furthermore, what if the doctor had taken an experiment to exclude the possibility of disease C? A dynamic logic will be proposed to deal with all these interesting issues.

The paper is organized as follows. After introducing a new formal structure representing evidence in Sect. 8.2, in Sect. 8.3 we define evidence models based on this structure rather than the evidence sets in van Benthem and Pacuit (2011a). We also discuss how to define knowledge in this model. To elucidate the new definition of knowledge, we compare it with the notion of safe belief. Furthermore, we look

[1] In general, our approach of defining knowledge can be embedded in the tradition of evidentialism (cf. Conee and Feldman 2004), where evidence is the most fundamental concept. Contrary to this view, Williomson (1997) interprets evidence in terms of knowledge. This paper contributes little to this debate about the conceptual hierarchy, assuming only the priority of evidence.

at belief and knowledge update caused by evidence dynamics, introduce a new evidence dynamic modality, and find the recursion axioms for it in Sect. 8.4. In the final section we conclude and point out some directions for future work.

8.2 Evidence as Sets of Hypotheses

According to Example 8.1.1, for knowledge attribution, it is essential to make clear what hypotheses the agent's evidence can support, besides the agent's "diagnosis" by her evidence. Therefore, we represent one piece of evidence by a set of possible hypotheses supported by that piece of evidence: symptom S can be represented as $\{A\}$ and symptom T can be represented as $\{B, C\}$.

Generally, we take a hypothesis as a proposition, which means we can characterize a hypothesis as a set of possible worlds in certain logical space W. Then one piece of evidence can be $H \subseteq \wp(W)$ (call it a hypothesis set). We say a piece of evidence H supports certain hypothesis h if $h \in H$.

As there is no evidence supporting a contradiction and everything as evidence should always support something (at least evidence itself), the empty set should not be in any hypothesis set ($\emptyset \notin H$) and hypothesis set itself should not be empty set ($H \neq \emptyset$).

On the other hand, when one piece of evidence supports two hypotheses, just as symptom T supports diseases B and C in Example 8.1.1, it is reasonable to expect that the disjunction of these two hypotheses should also be supported by that piece of evidence, just like B$\vee$C can be one hypothesis supported by symptom T. But we would hesitate to admit that "the patient got disease B or it is raining now in Beijing" is supported by symptom T directly, because the second disjunct "it is raining now in Beijing" seems totally irrelevant to symptom T.

So, in other words, it is not unreasonable to require that

$$\forall h, h' \in H : h \cup h' \in H.$$

but not

$$\forall h \in H : h \subseteq h' \Rightarrow h' \in H.$$

We denote the set of all hypothesis sets satisfying the three conditions as $\mathbb{H}$:

Definition 8.2.1. Let $\mathbb{H} \subseteq \wp\wp(W)$. $H \in \mathbb{H}$ if and only if H satisfies the following three conditions:

- $\emptyset \notin H$
- $H \neq \emptyset$
- $\forall h, h' \in H : h \cup h' \in H$.

This characterization of evidence is far from complete. It does not consider the agent's information concluded based on her evidence. In the next section, we take

the agent's "diagnosis" into account and characterize the evidential state of an agent, based on which we will then define the agent's belief and knowledge.

8.3 Belief, Reliability and Knowledge

8.3.1 Evidence Pair

We attach to each "symptom" the "diagnosis" the agent makes by this "symptom" to form an evidence pair (I, H), where $I \subseteq W$ formally captures the agent's "diagnosis", i.e. the information the agent gets from evidence $H \in \mathbb{H}$. A thorough grasp of the relationship between I and H requires a deep look into the process of making "diagnosis", which involves a lot of reasoning activities, e.g. abduction, induction and so on. We will not address this problem in this paper, but it should be an open problem for our future work.

Let $EP = \{(I, H) \mid I \subseteq W \text{ and } H \in \mathbb{H}\}$, we construct the following EP model:

Definition 8.3.1 (EP model). An **EP model** is a tuple $\mathcal{M} = \langle W, E, V\rangle$ with W a non-empty set of worlds, $E \subseteq W \times \wp(EP)$ an evidence relation, and $V : \mathbf{At} \to \wp(W)$ a valuation function. $E(w)$ is written for the set $\{(I, H) \mid wE(I, H)\}$, representing the agent's body of evidence at w. Four constraints are imposed on the evidence sets:

- For each $w \in W$ and each hypothesis set $H \in \mathbb{H}$, $(\emptyset, H) \notin E(w)$ (the agent can never get contradictory information from any piece of evidence);
- For each $w \in W$, $(W, \{W\}) \in E(w)$ (the agent knows his space);
- For each $w \in W$, if $(I, H), (I', H') \in E(w)$ and $H = H'$, then $I = I'$ (from the same pieces of evidence the agent gets the same information);
- For each w, $E(w)$ is a finite set (the pieces of evidence the agent possesses are finite).

Because the first element I of an evidence tuple (I, H) is essentially the same as the evidence set in van Benthem and Pacuit (2011b), we follow van Benthem and Pacuit (2011b)'s way of defining belief: the agent can aggregate all her information she has evidence for to form a consistent belief.

Notice the agent's belief cannot be defined by simply taking the intersection of all the $I \in \{X \subseteq W \mid (X, H) \in E(w)\}$, for $E(w)$ does not necessarily satisfy that $\bigcap\{I \subseteq W \mid (I, H) \in E(w)\} \neq \emptyset$. Despite of that, the agent can still combine all the mutually compatible evidence-based information sets which will not lead to an empty intersection:

Definition 8.3.2. A w-**scenario** is a maximal collection $\mathcal{X} \subseteq \{I \mid (I, H) \in E(w)\}$ that has the FIP, i.e., the finite intersection property: for each finite subfamily $\{X_1, \ldots, X_n\} \subseteq \mathcal{X}, \bigcap_{1 \leq i \leq n} X_i \neq \emptyset$.

Here we define an atomic sentence as a set of possible worlds in an EP model, then $\neg P = W \setminus P$, $P \wedge Q = P \cap Q$ and other boolean connectives can be defined in terms of $\neg$ and $\wedge$. Then the agent's belief at w in an EP model $\mathcal{M}$ can be defined upon these w-**scenario**:

Definition 8.3.3 (Belief).

– $\mathcal{M}, w \models BP$ iff for each w-scenario $\mathcal{X}$, $\bigcap \mathcal{X} \subseteq P$

The only difference between our concept of belief and that in van Benthem and Pacuit (2011b) is that in class of *EP* models, $\neg B\bot$ is valid, however, this is not the case for the class of evidence model in van Benthem and Pacuit (2011b), where $E(w)$ can be infinite.

Notice that the second element in the evidence pair (I, H) does not play any role in the definition of belief. However, when it comes to knowledge, it will do. It decides the reliability of the information obtained by the agent.

8.3.2 Knowledge as Reliable Belief

With belief defined in *EP* models, it is natural to ask about the truth condition of knowledge operator. We propose to understand knowledge as reliable belief, where belief is defined as above and reliability is ensured by the agent's evidence. Some belief is reliable if and only if it can be entailed by combining all the information which the agent has reliable evidence for. In Example 8.1.1 it is obvious that symptom S is reliable evidence for disease A but symptom T is not reliable evidence for disease B, because symptom T also indicates the possibility of disease C and it cannot be concluded only by symptom T which disease should be the right one. However, it can be concluded from symptom T that "the patient got either disease B or disease C". So if the doctor makes some diagnosis logically weaker than this hypothesis and also supported by symptom T, then we can say that the doctor has reliable evidence for her diagnosis.

Let $H_{min} = \{h \in H \mid \neg\exists h' \in H : h' \subset h\}$, the agent has reliable evidence for her "diagnosis" if and only if

$$(I, H) \in E(w) \text{ and } w \in \bigcup H_{min} \subseteq I$$

where $w \in I$ ensures the truth of I, $I \in H$ ensures I supported by some evidence, and $\bigcup H_{min} \subseteq I$ ensures that it can be concluded from evidence H that I is the case.

In order to express reliability of information obtained by the agent, we introduce a new operator $\bigcirc P$, which reads as "the agent has reliable information P". The truth condition of this operator is as follows:

Definition 8.3.4 (Reliable information). Given an EP model $\mathcal{M} = \langle W, E, V\rangle$ with $w \in W$,

– $\mathcal{M}, w \models \bigcirc P$ iff $\bigcap\{I \subseteq W \mid (I, H) \in E(w) \text{ and } w \in \bigcup H_{min} \subseteq I\} \subseteq P$

This truth condition says that the agent has reliable evidence for P if and only if P can be entailed by combining all the information which the agent has reliable evidence for.

Knowledge then can be expressed in *EP* models as follows:

Definition 8.3.5 (Knowledge).

– $KP := BP \wedge \bigcirc P$

Here $\bigcirc P$ ensures reliability of information P. So this notion of knowledge implies truth. On the other hand, it is also closed under implication because B and $\bigcirc$ are both closed under implication.

Fact 8.3.1. *In the class of EP models,*

– $\models KP \rightarrow P$
– $\models KP \rightarrow (K(P \rightarrow Q) \rightarrow KQ)$

In the above analysis, we have stressed that it is reliability that qualifies belief as knowledge, which is quite different from another tradition in epistemology which emphasizes robustness of belief (Lehrer and Paxson 1969; Swain 1974). In the next part, we compare these two notions, reliability and robustness of belief, and show that robustness implies reliability, but not vice versa.

8.3.3 Reliability and Robustness of Belief

Safe belief is the belief which the agent would not give up no matter what new true information she learned, which is proposed in Baltag and Smets (2008). And this notion actually coincides with the defeasibility analysis of knowledge formalized in Stalnaker (2006), which also embodies the robustness of belief under new information. However, Stalnaker (2006) points out the problem of this notion of knowledge, which "threatens to let any false belief to defeat too much of our knowledge" (Stalnaker 2006, p.190). In this section, we show that our notion of knowledge is a weaker notion than safe belief, and can avoid the problem of taking safe belief as knowledge.

We claim that safe belief can be defined in *EP* models as follows:

Definition 8.3.6 (Safe belief).

– $\mathcal{M}, w \models \Box Q$ iff $\bigcap\{I \subseteq W \mid (I, H) \in E(w) \text{ and } w \in I\} \cup \bigcup\{I \subseteq W \mid (I, H) \in E(w) \text{ and } w \notin I\} \subseteq Q$

To understand this truth condition of safe belief in *EP* models, we introduce a new operator expressing conditional belief – $B^P Q$, i.e. the agent would believe Q after learning P, which is essentially the same as that proposed in van Benthem and Pacuit (2011a).

In *EP* models, the agent's belief depends on his body of evidence. To express conditional belief, we have to consider the changes of the evidence-based information sets according to new information, which can be characterized by the following definition:

Definition 8.3.7. Suppose that $X \subseteq W$. Given a collection $\mathcal{X}$ of subsets of W, the relativization of $\mathcal{X}$ to X is the set $\mathcal{X}^X = \{Y \cap X \mid Y \in \mathcal{X}\}$. We say that a collection $\mathcal{X}$ of subsets of W has the finite intersection property relative to X (X-FIP) if for

each $\{X_1, \ldots, X_n\} \subseteq \mathcal{X}^X, \bigcap_{1 \leq i \leq n} X_i \neq \emptyset$. We say that $\mathcal{X}$ has the maximal X-FIP if $\mathcal{X}$ has X-FIP and no proper extension $\mathcal{X}'$ of $\mathcal{X}$ has the X-FIP.

Based on this definition, conditional belief $B^P Q$ is defined as follows:

– $\mathcal{M}, w \models B^P Q$ iff for each maximal P-FIP $\mathcal{X} \subseteq \{I \subseteq W \mid (I, H) \in E(w)\}$, $\bigcap \mathcal{X}^P \subseteq Q$

Then we can show why safe belief is defined as such in *EP* models by the following fact[2]:

Fact 8.3.2. *In any EP model $\mathcal{M}$ and any $w \in \mathcal{M}$: $\bigcap\{I \mid (I, H) \in E(w)$ and $w \in I\} \cup \bigcup\{I \mid (I, H) \in E(w)$ and $w \notin I\} \subseteq Q$ iff $\mathcal{M}, w \models B^P Q$ for any P such that $\mathcal{M}, w \models Q$.*

where "$\mathcal{M}, w \models B^P Q$ for any P such that $\mathcal{M}, w \models P$" is the exact interpretation of safe belief.

However, our modality of safe belief defined in *EP* models does not satisfy the laws of S4.3, which is a complete modal logic of this modality as introduced in Baltag and Smets (2008). Actually, positive introspection and .3 fail for our modality of safe belief, while the principles of **K** and **T** hold for it. In addition, the modalities of safe belief and knowledge defined in EP models satisfy the following relationship:

Fact 8.3.3. *In the class of EP models, we have the following validity:*

$$\models \Box P \rightarrow KP$$

This fact shows that robustness of belief implies reliability of belief in *EP* models. To prove this fact, we just need observe the facts that $\bigcap\{I \subseteq W \mid (I, H) \in E(w)$ and $w \in \bigcup H_{min} \subseteq I \in H\} \subseteq \bigcap\{I \subseteq W \mid (I, H) \in E(w)$ and $w \in I\}$; for any w-scenario $\mathcal{X}$ such that $\bigcap \mathcal{X} \cap \bigcap\{I \subseteq W \mid (I, H) \in E(w)$ and $w \in I\} \neq \emptyset$, $\bigcap \mathcal{X} \subseteq \bigcap\{I \subseteq W \mid (I, H) \in E(w)$ and $w \in I\}$, and for any w-scenario $\mathcal{X}$ such that $\bigcap \mathcal{X} \cap \bigcap\{I \subseteq W \mid (I, H) \in E(w)$ and $w \in I\} = \emptyset$, $\bigcap \mathcal{X} \subseteq \bigcup\{I \subseteq W \mid (I, H) \in E(w)$ and $w \notin I\}$.

Then why do we take knowledge as reliable belief but not safe belief?

Example 8.3.1. Ding has always been seeing his colleague Chang drive a Ford to work, and once Chang showed his car ownership papers to Ding. At the same time, there is another guy called Kai who also owns a Ford, but always rides to work instead of driving. And once he lied to Ding that he had no car. Therefore, Ding thinks that Chang owns the Ford but Kai does not.

Assume that Chang does own the Ford (p), Ding has no any other evidence against p and no any other evidence against "Kai owns no Ford" (q). In this situation, if someone told Ding that only one of these two propositions (p, q) is true, then Ding would hesitate about the truth of p given his acceptance of this new information.

[2]The direct proof of this fact can be found in Shi (2014). It can also be proved by transforming the evidence model to plausibility model, cf. Section 4.5 of van Benthem and Pacuit (2011b).

Let $W = \{pq, p\overline{q}, \overline{p}q, \overline{pq}\}, E(p\overline{q}) = \{(P, \{P\}), (Q, \{Q\})\}$ where pq means the world where p and q are both true, and $P = \{pq, p\overline{q}\}, Q = \{pq, \overline{p}q\}$, and $p\overline{q}$ is the actual world. It is easy to check that $\mathcal{M}, p\overline{q} \models \neg B^{(P\wedge\neg Q)\vee(\neg P\wedge Q)}P$ and so $\mathcal{M}, p\overline{q} \models \neg\Box p$. However, it seems too demanding to deny that Ding knew p just because he holds an irrelevant false belief of q. If knowledge necessarily implies safe belief, then the agent can know little. Just as the truth condition of safe belief in EP model itself shows, too many evidence-based information sets are taken into account even if they may be totally irrelevant.

To conclude this section, let us remind the readers what Stalnaker once said, "Perhaps the explanation of epistemic accessibility, in the case where conditions are not fully normal, and not all of the agent's beliefs are true, should focus more on the causal sources of beliefs, rather than on how agents would respond to information that they do not in fact receive" (Stalnaker 2006, p. 192). Our proposed definition of evidence-based knowledge emphasizes the causal sources of beliefs, and as well, the reliability of the support relation between the causal sources and information got from them. This may be seen as a contribution to Stalnaker's remarks. Just as shown in the example, it holds that $\mathcal{M}, p\overline{q} \models KP$, which fits our intuition perfectly well.

8.4 Knowledge Update Caused by Evidence Dynamics

All the above analysis shows that the agent's knowledge hinges on her body of evidence. Therefore, once the agent's body of evidence is affected by new incoming information or evidence, her knowledge accordingly updates. In Introduction, we have mentioned two actions of evidence dynamics: "diagnosis" revision and evidence addition. We can also think of actions like evidence removal and evidence combination. Actually, in van Benthem and Pacuit (2011a) where one piece of evidence is taken simply as a set of possible worlds, the three actions, evidence addition, removal and modification, have been studied separately. In addition, belief update caused by these three actions of evidence is also studied systematically in van Benthem and Pacuit (2011a).

However, evidence sets in van Benthem and Pacuit (2011a) often cannot distinguish two different pieces of evidence, because the agent can have different evidence for the same information. This can be shown easily by our evidence pairs where the first element can be taken as one evidence set. Assume there are two different evidence pairs (I_1, H_1) and (I_2, H_2) where $I_1 = I_2$ and $H_1 \neq H_2$. If we represent them by two evidence sets, then the first one is I_1 and the second one is I_2, which are indistinguishable. Therefore, some of the subtleties in evidence dynamics are ignored, namely H which affects knowledge attribution. On the other hand, let $I_1 \neq I_2, H_1 = H_2$ and assume the agent has had evidence (I_1, H_2). Then the addition of (I_2, H_2) seems not a simple evidence addition. It consists of two actions, removal of (I_1, H_2) and addition of (I_2, H_2), just like the diagnosis revision we mentioned in Introduction. When someone reminds the doctor of the possibility of disease C, then the doctor's diagnosis by symptom T will update to "the patient got disease B or C". Evidence addition happens only when the new piece of evidence has not been in the agent's body of evidence.

In this section we introduce a dynamic operator describing evidence dynamics where evidence is taken as evidence pairs instead of evidence sets. Furthermore, we focus on the belief and knowledge update under this evidence dynamics.

8.4.1 Evidence Acceptance $[*(P, \{P_1, \ldots, P_n\})]$

The following model transformation represents the process of adding an evidence pair:

Definition 8.4.1. Let $\mathcal{M} = \langle W, E, V\rangle$ be an EP model. The model $\mathcal{M}^{*(P,\{P_1,\ldots,P_n\})} = \langle W^{*(P,\{P_1,\ldots,P_n\})}, E^{*(P,\{P_1,\ldots,P_n\})}, V^{*(P,\{P_1,\ldots,P_n\})}\rangle$ where

- $W^{*(P,\{P_1,\ldots,P_n\})} = W$
- $V^{*(P,\{P_1,\ldots,P_n\})} = V$
- $\forall w \in W : E^{*(P,\{P_1,\ldots,P_n\})}(w) = (E(w) \setminus \{(X, \{P_1, \ldots, P_n\}) \mid (X, \{P_1, \ldots, P_n\}) \in E(w)\}) \cup \{(P, \{P_1, \ldots, P_n\}) \mid P \neq \emptyset\}$

Evidence removal can essentially be an addition of (I, H) where $I = \emptyset$. And more interestingly, it shows that the addition of an evidence pair is not simply evidence addition or removal, but a combination of these two actions.

It can be described by a dynamic operator $[*(P, \{P_1, \ldots, P_n\})]Q$, stating "after accepting the new piece of evidence $\{P_1, \ldots, P_n\}$ and information P got from it, Q is the case":

- $\mathcal{M}, w \models [*(P, \{P_1, \ldots, P_n\})]Q$ iff $\mathcal{M}^{*(P,\{P_1,\ldots,P_n\})}, w \models EP_1 \wedge \ldots \wedge EP_n$ implies $\mathcal{M}^{*(P,\{P_1,\ldots,P_n\})}, w \models Q$

where E is an existence operator. The precondition is that each P_i is true at some possible worlds, since we have stipulated that no hypothesis can be contradictory.

Then what effect does this operation have on belief and knowledge in the *EP* models? As usual, we attempt to find the recursion axioms which describe the epistemic change before and after the action takes place.

8.4.2 Recursion Axioms

For this purpose, we introduce two static modalities $B^{(P,\{P_1,\ldots,P_n\})}Q$ and $\bigcirc^{(P,\{P_1,\ldots,P_n\})}Q$. Let $E'(w) = (E(w) \setminus \{(X, \{P_1, \ldots, P_n\}) \mid (X, \{P_1, \ldots, P_n\}) \in E(w)\}) \cup \{(P, \{P_1, \ldots, P_n\})\}$, we have their truth conditions in the EP model:

- $\mathcal{M}, w \models B^{(P,\{P_1,\ldots,P_n\})}Q$ iff for each maximal collection with FIP $\mathcal{X} \subseteq \{X \subseteq W \mid (X, H) \in E'(w)\}$, $\bigcap \mathcal{X} \subseteq Q$
- $\mathcal{M}, w \models \bigcirc^{(P,\{P_1,\ldots,P_n\})}Q$ iff $\bigcap\{I \subseteq W \mid (I, H) \in E'(w)$ and $w \in \bigcap H_{min} \subseteq I \in H\} \subseteq Q$

$B^{(P,\{P_1,\ldots,P_n\})}Q$ can be seen as conditional belief except that it is conditional on evidence instead of direct information.

It seems that the operation of $(\cdot)^{(P,\{P_1,\ldots,P_n\})}$ on EP models is the same as that of $[*(P,\{P_1,\ldots,P_n\})]$. In fact, they are quite different. The former is simply a local operation but the latter is a universal operation on every state.

With these two new modalities, we can form the recursion axioms for belief and knowledge after evidence acceptance:

$$[*(P,\{P_1,\ldots,P_n\})]BQ \leftrightarrow ((EP_1 \wedge \ldots \wedge EP_n) \rightarrow B^{(P,\{P_1,\ldots,P_n\})}[*(P,\{P_1,\ldots,P_n\})]Q)$$

$$[*(P,\{P_1,\ldots,P_n\})] \bigcirc Q \leftrightarrow ((EP_1 \wedge \ldots \wedge EP_n) \rightarrow \bigcirc^{(P,\{P_1,\ldots,P_n\})}[*(P,\{P_1,\ldots,P_n\})]Q)$$

$$[*(P,\{P_1,\ldots,P_n\})]KQ \leftrightarrow [*(P,\{P_1,\ldots,P_n\})]BQ \wedge [*(P,\{P_1,\ldots,P_n\})] \bigcirc Q$$

8.4.3 Language Extension

Yet we are not done. To get a complete logic, we also need to find recursion axioms for the two new modalities, we first extend our basic language:

Definition 8.4.2. The language $\mathcal{L}_D$ of this dynamic logic is generated by the following grammar:

$$p \mid \neg\varphi \mid \varphi \wedge \varphi \mid B^{\overline{(\varphi,\{\varphi,\ldots,\varphi\})}}\varphi \mid \bigcirc^{\overline{(\varphi,\{\varphi,\ldots,\varphi\})}}\varphi \mid [*(\varphi,\{\varphi,\ldots,\varphi\})]\varphi \mid U\varphi$$

where $\overline{(\varphi,\{\varphi,\ldots,\varphi\})}$ is a finite sequence $(\varphi^1,\{\varphi_1^1,\ldots,\varphi_i^1\}),\ldots,(\varphi^m,\{\varphi_1^m,\ldots,\varphi_j^m\})$, and U is a universal operator whose dual is E.

The new modalities $B^{\overline{(\varphi,\{\varphi_1,\ldots,\varphi_n\})}}\psi$ and $\bigcirc^{\overline{(\varphi,\{\varphi_1,\ldots,\varphi_n\})}}\psi$ can be seen as a generalization of $B^{(\varphi,\{\varphi_1,\ldots,\varphi_n\})}\psi$ and $\bigcirc^{(\varphi,\{\varphi_1,\ldots,\varphi_n\})}\psi$. In this language, we can define $B\varphi$ and $\bigcirc\varphi$ as $B^{\emptyset}\varphi$ and $\bigcirc^{\emptyset}\varphi$, respectively.

Let $\overline{(\varphi,\{\varphi_1,\ldots,\varphi_n\})}_{max}$= $\{(\llbracket\varphi^i\rrbracket, \{\llbracket\varphi_1^i\rrbracket, \ldots, \llbracket\varphi_n^i\rrbracket\}) \mid \neg\ \exists\ j > i\colon \{\llbracket\varphi_1^j\rrbracket, \ldots, \llbracket\varphi_n^j\rrbracket\}= \{\llbracket\varphi_1^i\rrbracket,\ldots,\llbracket\varphi_n^i\rrbracket\}\}$ and $E''(w) = (E(w) \setminus \{(X,H) \in E(w) \mid (Y,H) \in \overline{(\varphi,\{\varphi_1,\ldots,\varphi_n\})}_{max}\}) \cup \overline{(\varphi,\{\varphi_1,\ldots,\varphi_n\})}_{max}$.

Definition 8.4.3. Given an EP evidence $\mathcal{M} = \langle W,E,V\rangle$ with $w \in W$ and φ in the language $\mathcal{L}_D$, define $\mathcal{M},w \models \varphi$ as follows:

- $\mathcal{M},w \models p$ iff $w \in V(p)$
- $\mathcal{M},w \models \neg\varphi$ iff $\mathcal{M},w \not\models \varphi$
- $\mathcal{M},w \models \varphi \wedge \psi$ iff $\mathcal{M},w \models \varphi$ and $\mathcal{M},w \models \psi$
- $\mathcal{M},w \models B^{\overline{(\varphi,\{\varphi_1,\ldots,\varphi_n\})}}\psi$ iff for each maximal collection with FIP $\mathcal{X} \subseteq \{X \subseteq W \mid (X,H) \in E''(w)\}$, $\bigcap \mathcal{X} \subseteq \llbracket\psi\rrbracket$
- $\mathcal{M},w \models \bigcirc^{\overline{(\varphi,\{\varphi_1,\ldots,\varphi_n\})}}\psi$ iff $\bigcap\{I \subseteq W \mid (I,H) \in E''(w)$ and $w \in \bigcup H_{min} \subseteq I \in H\} \subseteq \llbracket\psi\rrbracket$
- $\mathcal{M},w \models U\varphi$ iff $\forall v \in W : \mathcal{M},v \models \varphi$.

– $\mathcal{M}, w \models [*(\varphi, \{\varphi_1, \ldots, \varphi_n\})]\psi$ iff $\mathcal{M}^{*(\llbracket\varphi\rrbracket, \{\llbracket\varphi_1\rrbracket, \ldots, \llbracket\varphi_n\rrbracket\})}, w \models E\varphi_1 \wedge \ldots \wedge E\varphi_n$ implies $\mathcal{M}^{*(\llbracket\varphi\rrbracket, \{\llbracket\varphi_1\rrbracket, \ldots, \llbracket\varphi_n\rrbracket\})}, w \models \psi$

In the truth conditions, E'' reflects the operation of the sequence of evidence pairs on EP models, where only evidence pairs belonging to $\overline{(\varphi, \{\varphi_1, \ldots, \varphi_n\})}_{max}$ have an impact. This means that only the latest "diagnosis" the agent makes from a piece of evidence influences her evidential state. The acceptance of a new evidence pair means the agent forsakes her old "diagnosis" from the same piece of evidence and holds a new "diagnosis". The "diagnosis" corresponding to the latest appearance of a piece of evidence in the sequence dominates the "diagnosis" she made before from this piece of evidence.

With this generalization of the language, we can find complete recursion axioms for the dynamic modality:

Theorem 8.4.1. *The dynamic logic is axiomatized by (a) the static base logic of EP models for language $\mathcal{L}$ which does not include the dynamic modality; (b) the minimal modal logic for the evidence dynamic modality; and (c) the following set of recursion axioms:*

A1 $[*(\varphi, \{\varphi_1, \ldots, \varphi_m\})]p \leftrightarrow p$
A2 $[*(\varphi, \{\varphi_1, \ldots, \varphi_m\})]\neg\psi \leftrightarrow \neg[*(\varphi, \{\varphi_1, \ldots, \varphi_m\})]\psi$
A3 $[*(\varphi, \{\varphi_1, \ldots, \varphi_m\})](\psi \wedge \chi) \leftrightarrow [*(\varphi, \{\varphi_1, \ldots, \varphi_m\})]\psi \wedge [*(\varphi, \{\varphi_1, \ldots, \varphi_m\})]\chi$
A4 $[*(\varphi, \{\varphi_1, \ldots, \varphi_m\})]B^{\overline{(\psi, \{\psi_1, \ldots, \psi_n\})}}\chi \leftrightarrow$
$((E\varphi_1 \wedge \ldots \wedge E\varphi_m) \rightarrow B^{(\varphi, \{\varphi_1, \ldots, \varphi_m\}), \overline{(\psi, \{\psi_1, \ldots, \psi_n\})}}[*(\varphi, \{\varphi_1, \ldots, \varphi_m\})]\chi)$
A5 $[*(\varphi, \{\varphi_1, \ldots, \varphi_m\})] \bigcirc^{\overline{(\psi, \{\psi_1, \ldots, \psi_n\})}} \chi \leftrightarrow$
$((E\varphi_1 \wedge \ldots \wedge E\varphi_m) \rightarrow \bigcirc^{(\varphi, \{\varphi_1, \ldots, \varphi_m\}), \overline{(\psi, \{\psi_1, \ldots, \psi_n\})}}[*(\varphi, \{\varphi_1, \ldots, \varphi_m\})]\chi)$
A6 $[*(\varphi, \{\varphi_1, \ldots, \varphi_m\})]U\psi \leftrightarrow ((E\varphi_1 \wedge \ldots \wedge E\varphi_m) \rightarrow U[*(\varphi, \{\varphi_1, \ldots, \varphi_m\})]\psi$

The validity of A4 and A5 can be proved by observing that the range of possible worlds in $\mathcal{M}^{(\varphi, \{\varphi_1, \ldots, \varphi_m\})}$ where χ is evaluated is the same as that in $\mathcal{M}$ where $[*(\varphi, \{\varphi_1, \ldots, \varphi_m\})]\chi$ is evaluated, as $(E^{*(\varphi, \{\varphi_1, \ldots, \varphi_m\})}(w) \setminus \{(X, H) \in E^{*(\varphi, \{\varphi_1, \ldots, \varphi_m\})}(w) \mid (Y, H) \in \overline{(\psi, \{\psi_1, \ldots, \psi_n\})}_{max}\}) \cup \overline{(\psi, \{\psi_1, \ldots, \psi_n\})}_{max}$ equals to $(E(w) \setminus \{(X, H) \in E(w) \mid (Y, H) \in \overline{((\varphi, \{\varphi_1, \ldots, \varphi_m\}), (\psi, \{\psi_1, \ldots, \psi_n\}))}_{max}\}) \cup \overline{((\psi, \{\psi_1, \ldots, \psi_m\}), (\psi, \{\psi_1, \ldots, \psi_n\}))}_{max}$.

The recursion axioms essentially reflect the same effect of two operations $[*(\varphi, \{\varphi_1, \ldots, \varphi_m\})]$ and $(\cdot)^{(\varphi, \{\varphi_1, \ldots, \varphi_m\})}$ on $E(w)$. The only difference between them is that $[*(\varphi, \{\varphi_1, \ldots, \varphi_m\})]$ also changes $E(v)$ for $v \neq w$, but $(\cdot)^{(\varphi, \{\varphi_1, \ldots, \varphi_m\})}$ has no impact on other possible worlds. Therefore, we have the following validity in the class of *EP* models:

Fact 8.4.1. *In the class of EP models, for any formula $\chi \in \mathcal{L}_D$ without any modality in it,*

– $\models [*(\varphi, \{\varphi_1, \ldots, \varphi_m\})]B^{\overline{(\psi, \{\psi_1, \ldots, \psi_m\})}}\chi \leftrightarrow$
$((E\varphi_1 \wedge \ldots \wedge E\varphi_m) \rightarrow B^{(\varphi, \{\varphi_1, \ldots, \varphi_m\}), \overline{(\psi, \{\psi_1, \ldots, \psi_n\})}}\chi)$
– $\models [*(\varphi, \{\varphi_1, \ldots, \varphi_m\})] \bigcirc^{\overline{(\psi, \{\psi_1, \ldots, \psi_n\})}} \chi \leftrightarrow$
$((E\varphi_1 \wedge \ldots \wedge E\varphi_m) \rightarrow \bigcirc^{(\varphi, \{\varphi_1, \ldots, \varphi_m\}), \overline{(\psi, \{\psi_1, \ldots, \psi_n\})}}\chi)$

This shows that the difference between belief (reliable information) after the acceptance of new evidence and belief (reliable information) conditional on the same evidence only appears in higher-order epistemic attitude.

8.5 Conclusion and Future work

This paper has followed van Benthem and Pacuit (2011a)'s characterization of evidence and developed its ideas, more specifically, we have taken the evidential support relation into account, which has been generally considered at the syntactic level in the tradition of justification logic. The reliability of evidential support relation has been our core concern in this context. Technically, modelling evidence as evidence pair makes it possible to express reliability in our framework. Then we formalized our new idea of defining knowledge based on reliable evidence and belief. With the assistance of our formal machinery, several philosophical issues have become clear, in particular, the relationship between notions of belief, evidence and knowledge, as well as the properties of robustness and reliability of belief. Finally, we explored the knowledge (belief) update caused by evidence dynamics, this has naturally extended our static language. Interestingly, this has also brought us a new concept – belief conditional on new evidence.

Nevertheless, further issues arise, as listed below. We will leave them for our future work:

More Specific Structure of the Agent's Evidential State The formalization of the agent's body of evidence in this paper is still simple. We did not consider the plausibility order of each piece of evidence and the relevance between different pieces of evidence. A natural direction for future work is to introduce the plausibility order into the agent's body of evidence and to group the relevant pieces of evidence together so that the agent considers only most plausible evidence and combines only the evidence in the same group. The group of relevant pieces of evidence may be decided by questions or topics. This will connect to the existing formal analysis of same issues in epistemic logic (van Benthem and Minic 2012; Ciardelli et al. 2013).

Evidence Pairs Again The current set up of the evidence pair is rather abstract. The relation between I and H is worth further consideration. For instance, how can the agent get I from certain piece of evidence H? The possible answer may involve the agent's background information and ability of inference. And what conditions should one piece of evidence satisfy to support an hypothesis? Answering these questions would lead to a more complete and systematic theory about evidence and knowledge.

Multi-agent Scenarios Our perspective in this paper has been mostly one agent's view, and we did not consider the multi-agent scenarios. However, the problem of forming an agent's belief and knowledge by aggregating evidence-based

information sets is very relevant to the discussion on the formation of social belief (knowledge) by aggregating other individual's beliefs (Liu et al. 2014). Adding the social aspect to the present framework will definitely open a new arena for us.

References

Artemov, S.: The logic of justification. Rev. Symb. Log. **1**, 477–513 (2008)
Baltag, A., Smets, S.: A qualitative theory of dynamic interactive belief revision. Texts Log. Games **3**, 9–58 (2008)
Baltag, A., Renne, B., Smets, S.: The logic of justified belief, explicit knowledge, and conclusive evidence. Ann. Pure Appl. Log. **165**, 49–81 (2014)
Ciardelli, I., Groenendijk, J., Roelofsen, F.: Inquisitive semantics: a new notion of meaning. Lang. Linguist. Compass **7**(9), 459–476 (2013)
Conee, E., Feldman, R.: Evidentialism: Essays in Epistemology. Oxford University Press, New York (2004)
Dougherty, T. (ed.): Evidentialism and ts Discontents. Oxford University Press, New York (2011)
Lehrer, K., Paxson, T.: Knowledge: undefeated justified true belief. J. Philos. **66**(8), 225–237 (1969)
Liu, F., Seligman, J., Girard, P.: Logical dynamics of belief change in the community. Synthese **191**(11), 2403–2431 (2014)
Shi, C.: Logics of Evidence-Based Belief and Knowledge. Master's thesis, Tsinghua University (2014)
Stalnaker, R.: On logics of knowledge and belief. Philos. Stud. **128**(1), 169–199 (2006)
Swain, M.: Epistemic defeasibility. Am. Philos. Q. **11**(1), 15–25 (1974)
van Benthem, J., Minic, S.: Toward a dynamic logic of questions. J. Philos. Log. **41**(4), 633–669 (2012). http://dx.doi.org/10.1007/s10992-012-9233-7
van Benthem, J., Pacuit, E.: Dynamic logics of evidence-based belief. Stud. Log. **99**, 61–92 (2011a)
van Benthem, J., Pacuit, E.: Dynamic logics of evidence-based belief, Technical report, Institute for Logic, Language and Computation (2011b)
Williomson, T.: Knowledge as evidence. Mind **106**, 717–742 (1997)

Chapter 9
Public Announcements and Inconsistencies: For a Paraconsistent Topological Model

Can Başkent

Abstract In this paper, we discuss public announcement logic in topological context. Then, as an interesting application, we consider public announcement logic in a paraconsistent topological model.

Keywords Public announcement logic • Topological semantics • Homotopy • Paraconsistent logic

9.1 Introduction

9.1.1 Motivation

Public announcement logic is a formal framework that strives to express various dynamic aspects of knowledge change. Considered a kind of dynamic epistemic logic, public announcement logic works as follows. An external agent makes a truthful and public announcement, then the agents update their epistemic states by eliminating the possible worlds that do not agree with the announcement. For example, you may think that today is either Tuesday or Wednesday, then on TV you hear that it is actually Tuesday today. Then, you eliminate the possibility that today is Wednesday and come to know that today is Tuesday. Thus, after an announcement, you come to know the announcement.

Traditionally, public announcement logic (PAL, henceforth) adopts Kripke semantics (Plaza 1989; Gerbrandy 1999). Kripke frames and semantics enjoy a simplistic approach to modal logics in general, and makes it quite feasible to express various epistemic issues. However, Kripke semantics is not the only way to express truth in public announcement logic. In a relatively recent work, a topological semantics for public announcement logic was given (Başkent 2012). In that paper, the completeness and decidability results of PAL with respect to the topological semantics in several multi-agent frameworks were proven. Furthermore,

C. Başkent (✉)
Department of Computer Science, University of Bath, Bath, England
e-mail: can@canbaskent.net; www.canbaskent.net/logic

J. Redmond et al. (eds.), *Epistemology, Knowledge and the Impact of Interaction*,
Logic, Epistemology, and the Unity of Science 38, DOI 10.1007/978-3-319-26506-3_9

it was shown that topological semantics changes some aspects of PAL compared to Kripke semantics. For example, announcements may stabilize in more than ω steps in topological models, which cannot be the case in Kripke models. Moreover, topological models exhibit some unexpected properties when it comes to formal analysis of rationality and backward induction. In topological game models, where we consider a topology based on a game tree, under the assumption of rationality, the backward induction procedure can take more than ω steps (ibid).

In this work, we extend such results by focusing on the relation between topologies, public announcements and inconsistency-friendly logics, particularly paraconsistent logic. By paraconsistent logic, we mean the logical systems in which the explosion principle (which says that from a contradiction, everything follows) fails. Therefore, in paraconsistent systems, there are some formulas that do *not* follow from a contradiction. Paraconsistent logics help us build inconsistent but non-trivial theories. As we shall make it clear in due course, from an epistemological perspective, paraconsistency and dynamic epistemology show an appealing interaction. If the given universe admits ontological contradictions (namely, if *some* things *are* and *are not* at the same time), how can knowledge and the dynamic *change* of knowledge be expressed logically? How do they interact? What kind of dynamic semantics do we need, if we want a universal framework that can work with some adjustments both in classical and non-classical (paraconsistent, and also intuitionistic) structures?

One of the main motivation of this work comes from *impossible worlds* – worlds which satisfy contradictions. Adopting a model that admits some impossible worlds immediately raises some questions about the possibility of expressing dynamic epistemologies in such a model. That is what we achieve in this paper.

The organization of the current work is as follows. First, we briefly remind the reader the basic topological concepts and structures which we will need throughout the paper. Then, from a rather technical perspective, we will show that topological models indeed present a rich and wide variety of possibilities of mathematical modeling of dynamic epistemologies. Next, we will present paraconsistent public announcement logic with some examples.

9.1.2 Basics

Let us now start with some basic definitions to make this work more self contained. Here, we define the classical PAL with topological semantics following (Başkent 2012).

Given a non-empty set S, a topology σ is defined as a collection of subsets of S satisfying the following conditions.

- The empty set and S are in σ,
- The collection σ is closed under finite intersections and arbitrary unions.

We call the tuple (S, σ) a *topological space*. The members of the topology is called *opens*. Complement of an open set (with respect to the classical set theoretical complement) is called a *closed set*. A function defined on a topological space is *continuous* if the inverse image of an open is an open; *open* if the image of an open is an open. A function is called *homeomorphism* if it is a continuous function between topological spaces with a continuous inverse. Homeomorphic spaces possess the same topological properties.

The above definition of topological space is given based on open sets. A dual definition can be given with closed sets as the primitives. In this case, for a given set S, we define the topology σ as a collection of subsets of S with the following condition.

- The empty set and S are in σ,
- The collection σ is closed under arbitrary intersections and finite unions.

We will refer to the topological spaces defined this way as *closed set topologies*. In this case, members of the topology will be closed sets. Notice that this is a dual definition for topological spaces.

Given a topological space, we can define a logical model. Let $M = (S, \sigma, v)$ be a *topological model* where (S, σ) is a topology and v is a valuation function assigning subsets of S to propositional variables. We denote the extension of φ in a model M with $|\varphi|^M$, and define it as follows $|\varphi|^M = \{s \in S : s, M \models \varphi\}$. When it is obvious, we will drop the superscript. Then, for an announcement φ, we define the *updated model* $M'_\varphi = (S', \sigma', v')$ as follows. Set $S' = S \cap |\varphi|$, $\sigma' = \{O \cap S' : O \in \sigma\}$, and $v' = v \cap S'$. Thus, in PAL, an announcement is made and the states that do not satisfy the announcement are eliminated in a way that preserves the topological structure. Also, the updated models are parametrized based on the extension of the announcement, in which the agents come to know the announcement in the updated model. Logically equivalent formulas, and even formulas that have the same extensions in the given original model produce the same updated model. Also, notice that the new topology σ', which we obtained by relativizing σ, is a familiar one, and is called *the induced topology*, and is indeed a topology (Başkent 2012).

The language of topological PAL includes the epistemic modality K and the public announcement modality $[\cdot]$, and they are defined recursively in the standard fashion based on a given set of propositional variables. We denote the dual of K as L, and define it as $\mathsf{L}\varphi := \neg\mathsf{K}\neg\varphi$ for a negation symbol $\neg$. For simplicity, we only give the single agent PAL here.

In a topology, for a given set, we have the *interior* operator Int and the *closure* operator Clo which return the largest open set contained in the given set, and the smallest closed set containing the given set respectively. The extensions of modal/epistemic formulas depend on such operators. We put $|\mathsf{K}\varphi| = \mathsf{Int}(|\varphi|)$. Dually, we have $|\mathsf{L}\varphi| = \mathsf{Clo}(|\varphi|)$. Intuitively, extension of a modal formula is the interior (or the closure) of the extension of the formula. It is important to note that in the classical case, epistemic modal operators necessarily produce topological entities. However, it is not necessary that $|p|$ for a propositional variable p will be open or closed, as it simply does not follow from the definition.

The semantics of propositional variables and Booleans are standard. Let us give the semantics of the modalities here. For simplicity, we give the semantics for single agent here, and refer the reader to Başkent (2012) for various multi-agent extensions that require some more topological operations.

$$M, s \models \mathsf{K}\varphi \quad \text{iff} \quad \exists O \in \sigma.(s \in O \wedge \forall s' \in O, M, s' \models \varphi)$$
$$M, s \models [\varphi]\psi \quad \text{iff} \quad M, s \models \varphi \text{ implies } M', s \models \psi$$

The semantics of topological models makes it clear why topological models can distinguish a variety of epistemic properties that Kripke models cannot (van Benthem and Sarenac 2004). The reason is that the topological semantics for the epistemic modality K has Σ_2 complexity as it is of the form $\exists\forall-$, while Kripkean semantics offers Π_1 complexity as it is of the form $\forall-$. Also, even it does not directly fall within the scope of this paper, topological models handle infinitary cases better.

PAL with classical topological semantics admits the following standard reduction axioms.

- $[\varphi]p \leftrightarrow (\varphi \rightarrow p)$
- $[\varphi]\neg\psi \leftrightarrow (\varphi \rightarrow \neg[\varphi]\psi)$
- $[\varphi]\psi \wedge \chi \leftrightarrow [\varphi]\psi \wedge [\varphi]\chi$
- $[\varphi]\mathsf{K}\psi \leftrightarrow (\varphi \rightarrow \mathsf{K}[\varphi]\psi)$

In PAL, the rules of derivation are normalization ($\vdash \varphi \therefore \vdash \Box\varphi$) and modus ponens. Then, we have the expected completeness and decidability results.

Theorem 9.1 (Başkent 2012). *PAL in topological models is complete and decidable.*

The topological semantics for modal logics has been proposed in early 1940s even before the well-known Kripke semantics (van Benthem and Bezhanishvili 2007; McKinsey and Tarski 1946, 1944). The literature on the subject has evolved rapidly with a wide range of applications in philosophy and computer science, including various pointers to non-classical logics (Mints 2000; Goodman 1981). Within the family of non-classical logics, in this paper, we consider paraconsistent logics. We already gave a proof-theoretical definition of paraconsistency which underlines the fact that much of the work on the subject is from a proof-theoretical perspective. Yet, the current paper focuses on the semantical aspects of paraconsistency. *Dialetheism* is the view that suggests that there are true contradictions. Hence, dialetheism can be seen as a semantical counterpart of paraconsistency. In order to prevent an inflation of terminology, we will use both terms interchangeably when no confusion arises.

Paraconsistent logics span a very broad field with applications in computer science, philosophy and mathematical logic (Carnielli et al. 2007; da Costa et al. 2007; Priest 2002, 2008). We need to underline it at the beginning that, in this work, paraconsistency does not refer to the meta-logical (such as set theoretical, topological or arithmetical) properties of the models. For that reason, our definitions, proof methods and meta-logic are classical, and paraconsistency occurs at an object

level. Within the pluralistic world of paraconsistent logic, this is indeed one of the methods to introduce non-trivial inconsistencies into models.

Next, we first discuss various topological results for the classical PAL to show the strength of the topological semantics and the richness of the applications it provides. Then, we will take an additional step and discuss PAL in inconsistent models.

9.2 Topological Announcements

9.2.1 Homotopic Announcements

One of the advantages of working with topological models is the fact that a variety of topological tools can be used within this framework to express a broad range of epistemic and model theoretical situations. In this section, we will observe various strengths of topological semantics for public announcements.

We define *functional representation* of announcements with respect to a topological model $M = (S, \sigma, v)$ as follows. For a public announcement φ, we say φ is "functionally representable in M" if there is an open and continuous function $f^M_\varphi : (S, \sigma) \mapsto (S', \sigma')$ where $M'_\varphi = (S', \sigma', v')$ is the updated model. We will drop the superscript or subscript when they are obvious. Notice that open or continuous functions deal with only open (or dually, closed) sets. However, the extensions of each and every formula in the language (such as the extensions of ground formulas) are not necessarily an open set. Therefore, open or continuous functions do not take such formulas into account. Nevertheless, in a model where each formula necessarily has an open (or equivalently closed) set extension, functional representation still works.

We observe that public announcements are special kind of functional representations.

Theorem 9.2. *Every public announcement is functionally representable.*

Proof. Given $M = (S, \sigma, v)$, construct $M'_\varphi = (S', \sigma', v')$ with respect to the public announcement φ. Then, for every open $O \in \sigma$ in M, assign $f(O) = O'$ where $O' = O \cap S'$ in σ' in M'. Here, notice that O' can be the empty set for some $O \in \sigma$ which is perfectly OK as f is not imposed to be an one-to-one function. We claim f functionally represents φ.

Note that modal formulas necessarily produce open (or dually closed) sets as their extensions, and they are taken care of by the given function f. However, we may still have Boolean formulas which do not have open or closed extensions in the model. However, notice that they do not violate functional representation as the definition of functional representation quantifies over open sets.

Now, since both, O and O' are open, so f is an open map. Take $U' \in \sigma'$. Since, $U' = U \cap S'$ for some $U \in \sigma$, the inverse of image of U' under f is U which is an open in σ showing that f is continuous.

Thus, we conclude that f functionally represents φ. □

The converse of the above theorem is not true in general. Not every open and continuous function represents an announcement as it may not respect the valuation in the model. Now, we can use functions to represent the relation between the given (original) epistemic model and the updated model. This is indeed another way to represent the *dynamic* aspects of knowledge change in topological models.

Corollary 9.1. *In PAL, given topological models and the updated topological models may not be homeomorphic in general.*

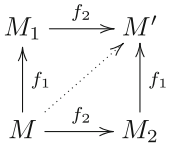
Proof. Functional representation of an announcement is not necessarily one-to-one, therefore may not be a homeomorphism. □

This is quite interesting. The above result indicates that not just knowledge may change after an announcement, but also the topological qualities of the model may alter. This is perhaps not surprising, as we would like the announcement to have an epistemic impact which may change some model theoretical properties of the model. This observation suggests the following definition.

Definition 9.1. Given two models $M = (S, \sigma, v)$ and $M' = (S'\sigma', v')$. We call M and M' *homeomorphic φ-models* if M' is the updated model of M with the public announcement φ, and there is a homeomorphism f from (S, σ) into (S', σ') that functionally represents φ.

Notice that homeomorphic model relation is not symmetric, but it is reflexive and transitive. Homeomorphic φ-models enjoy the same topological qualities after a specific public announcement (here, φ). In this context, arbitrary announcements (Balbiani et al. 2008) can be considered a generalization of homeomorphic φ-model to *homeomorphic models* that remain homeomorphic after *any* announcement.

For a given model M, consider two different announcements φ_1 and φ_2 representable by f_1 and f_2 respectively. Then, as $[\varphi_1][\varphi_2]\psi \leftrightarrow [\varphi_2][\varphi_1]\psi$, we have the following situation illustrated in the diagram.



For simplicity, we assume that M and M' are homeomorphic models. Then, what about the connection between M_1 and M_2? We can easily generalize this question to n many models. For public announcements φ_i functionally represented by f_i, and the updated models M_i obtained after announcing φ_i, one can ask about the relation between M_is? In order to give an answer to this question, we need homotopies.

Definition 9.2. Let S and S' be two topological spaces with continuous functions $f, f' : S \mapsto S'$. A homotopy between f and f' is a continuous function $H : S \times [0, 1] \mapsto S'$ such that for $s \in S$, $H(s, 0) = f(s)$ and $H(s, 1) = g(s)$.

The definition of homotopy can easily be extended to topological models. Given a topological model $M = (S, \sigma, v)$ we call the family of models $\{M_t = (S_t, \sigma_t, v_t)\}_{t\in[0,1]}$ generated by M and homotopic functions *homotopic models*. In the generation of valuation function v_ts of M_ts, we put $v_t = f_t(v)$. Homotopic models preserve truth, and they can be used to extend the definition of bisimulations in topological spaces (Başkent 2013).

Theorem 9.3. *Given M, consider a family of updated homeomorphic models $\{M_i\}_{i<\omega}$ each of which is obtained by an announcement φ_i representable by f_i. Then f_is are homotopic.*

Proof. Immediate. □

The converse of the above statement is not always true. Clearly, not each pair of updated models in a class of homotopic models can be obtained from one another by an update. Given M, consider the updated models M_1 and M_2 where the prior is obtained by an announcement of p while the latter $\neg p$. Even if there is a continuous transformation between M_1 and M_2, this transformation is not a public announcement.

Namely, there exists a smooth topological transformation from one updated model to another. Then, what is the epistemic meaning of it? Can we preserve truth under such a transformation?

We can make use of an earlier result here (Başkent 2013). Let $M = (S, \sigma, v)$ be a given model. Suppose $M_1 = (S_1, \sigma_1, v_1)$ and $M_2 = (S_2, \sigma_2, v_2)$ are updated models obtained after the announcements φ_1 and φ_2 respectively. Let the functions f_1 and f_2 represent φ_1 and φ_2 respectively. Then, there exists a homotopy $H : S \times [0, 1] \mapsto S$ such that $H(s, 0) = f(s)$ and $H(s, 1) = g(s)$ where $s \in S$. Now, observe that we also have $v_2 = f_2 f_1^{-1}(v_1)$. More importantly, we have another homotopy J such that $J(s, 0) = v_1$ and $J(s, 1) = v_2$. It is easy to notice that $J = v(H)$. Here, we discuss this example with only two updates, but the results can easily be generalized to n different updates.

In other words, the transformation between two updated models require a *renaming* or *restructuring* the real world.

Notice that homotopies discuss the topological connection between different announcements. The epistemic significance of this concept is the fact that now, at least in topological models, we can express how differentiated opinions can be transformed into each other under certain assumptions. This directly relates to *belief polarization* (Kelly 2008; Başkent et al. 2012). Thomas Kelly summarizes this phenomenon as follows.

> Suppose that two individuals – let us call them "You" and "I" – disagree about some nonstraightforward matter of fact. (...) Suppose next that the two of us are subsequently exposed to a relatively substantial body of evidence that bears on the disputed question. (...) What becomes of our initial disagreement once we are exposed to such evidence? (...)

> Exposure to evidence of a mixed character does not typically narrow the gap between those who hold opposed views at the outset. Indeed, worse still: not only is convergence typically not forthcoming, but in fact, exposure to such evidence tends to make initial disagreements even more pronounced. Kelly (2008)

This interesting, yet very common and basic phenomenon can easily be formalized in terms of public announcements. In this case, the announcement ("the substantial body of evidence") creates different updates on different agents. So far, this is perfectly normal. What is interesting is that the updated models of two agents are not transformable to each other – that is they are not homotopic. Thus, they are polarized.

In this case, homotopic models represent *degrees* of belief or knowledge where the models can be, step by step, translated to each other, and such a translation follows a topologically meaningful pattern – it preserves the topological and ideally (if it is a special kind of homotopy) model theoretical properties of the models in question. However, polarized beliefs and knowledge of two agents, in this case, cannot be transformed into each other, by the mere definition of polarization. Thus, they cannot be homotopic. This is a simple but direct application of homotopic public announcements.

In short, there is a close connection between various topological transformations and model updates after public announcements, and topological PAL models enjoy various techniques imported from pure topology. Moreover, they may correspond to various interesting epistemic concepts that are relevant for dynamic epistemic logic.

9.3 Paraconsistent Public Announcements

In classical logic, contradictions are never satisfied. However, in modal philosophical logic there is an interesting conceptual and philosophical notion, called *impossible worlds*. By *impossible worlds*, let us denote those states which satisfy some contradictions, define them as $\{x : x \models \varphi \wedge -\varphi \text{forsome} \varphi\}$ for a negation symbol $-$. Then, the natural question is how to epistemically update an epistemic model with impossible worlds.

For example, if we consider God as an impossible state in our mental model, how can we then update our mental epistemic model after we hear about a person healing the blind or splitting the Moon? Mental models may possess some contradictions, yet, they still function in a (relatively) rational and sound fashion. People believe in gods, they believe in miracles, yet they still function mostly rationally – both dynamically and epistemically. How can we portray such epistemic situations when an external announcement updates the models with impossible worlds?

Law, as a major platform for inconsistencies, exhibit similar puzzling situations.

> Suppose that there is a certain country which has a constitutional parliamentary system of government. And suppose that its constitution contains the following clauses. In a parliamentary election:

(1) no person of the female sex shall have the right to vote;
(2) all property holders shall have the right to vote.

(Priest 2006, p. 184)

Let us denote the above rules as public announcements φ_1, φ_2 respectively. Therefore, when the Law (1) was introduced, we can consider it as $[\varphi_1]$, and similarly Law (2) as $[\varphi_2]$. The introduction of new laws to the legal system can be thought of as public announcements. For simplicity, consider them as a simultaneous announcement of the form $[\varphi_1 \wedge \varphi_2]$. Therefore, when $[\varphi_1 \wedge \varphi_2]$ is announced, the states that satisfy the contradictory statement will be kept – which is the set of female property holders, in this example. This announcement does not (and should not) trivialize the system. In this case, contradictions exist, yet we are supposed to reason soundly in this model, we cannot let the model get trivialized or explode.

Another motivational example comes from a neighboring field of belief revision. Priest discusses AGM style belief revision from a paraconsistent perspective, and revises the AGM postulates (Priest 2001). Belief is defined weaker than knowledge. Therefore, the immediate next step is to consider knowledge in a paraconsistent universe, and observe how it changes.

Our goal now is to give a formal model which can descriptively and normatively express such situations.

9.3.1 Models

Topological semantics provides a versatile tool to express truth in a wide range of classical and non-classical logics. As we already showed, it is also a wise choice to express various dynamic and modal issues.

While discussing the classical topological semantics, we underlined that only the modal formulas produce topological sets (opens or closeds). Boolean formulas do not necessarily produce such sets. Let us now assume that we have a closed set topology where each member of the topology is a closed set, and stipulate further that the extensions of propositional variables are also closed sets. If propositional variables are closed sets, then their arbitrary intersections and finite unions will remain closed. Therefore, conjunctions and disjunctions of such propositional variables will still be closed sets. However, this stipulation makes an important difference for negation as the compliment of a closed set is not necessarily a closed set. For that reason, we cannot use the standard definition of negation as the set theoretical complement on the extension of the formula. So, we need to redefine it in closed set topologies. In our system, we define negation as the "closure of the complement" (Başkent 2013; Goodman 1981; Mortensen 2000). Let us denote this paraconsistent negation by $-$.

As an illustration, consider the formula $p \wedge -p$. Let us say that the extension of p is $O \in \sigma$ where σ is a closed set topology, and O is a closed set. Then the extension of $p \wedge -p$ is $O \cap \mathsf{Clo}(\overline{O})$ which is $\partial(O)$, where $\partial(\cdot)$ is the boundary operator which

is defined as $\partial O := \mathsf{Clo}(O) - \mathsf{Int}(O)$, and $\overline{O}$ denotes the (classical) set theoretical compliment of O. Therefore, the contradictions are satisfied at the boundary points. Thus, we now have a paraconsistent logic. The reason why explosion fails is because for some formula φ, the extension of $\varphi \wedge -\varphi$ is not necessarily an empty set, but it is $\partial(O)$ for some closed set O. Thus, it is not necessarily a subset of every set, so not every formula follows from a contradiction in this system, failing the explosion property.

However, we need to elaborate a bit more on the epistemological meaning of the use of paraconsistent spaces in the context of public announcement logic. The classical PAL heavily depends on the law of non-contradiction. An external and truthful announcement is made. Then, the agents update their epistemic models by eliminating the states in their model which do not agree with the announcement, followed by the reducing the epistemic accessibility relation or the topology and the valuation with respect to the new, updated model. Therefore, the classical PAL does not *control* the inconsistencies, it completely eliminates them. Yet, in paraconsistent spaces, some contradictions need not be eliminated as they do not trivialize the theory. In short, the main problem caused by inconsistencies is that they trivialize the theory due to the choice of the underlying logic. Therefore, if there exists some contradictions that do not trivialize the theory (again, due to the choice of the underlying logical framework), there seems to be no need to eliminate them. This is our pivotal point for paraconsistent PAL. Also, notice that intuitionistic logic also admits explosion, thus suffers from the same problem as the classical logic.

Here, notice that we do not focus on inconsistent announcements or non-truthful announcements per se. Our framework reflects paraconsistent modal realism, and allows inconsistent possible worlds. Moreover, we also follow the standard "state elimination based" paradigm for PAL – with some differences which will be clarified in due time. Model theoretically, we can also eliminate the accessibility relation arrows or relativize only the topology and leaving the universe intact and keep the states. From modal logical perspective, there seems to be no model theoretical difference between these methods.

In paraconsistent spaces, public announcements obtain a broader meaning. Namely, when φ is announced in a paraconsistent space, it simply means "Keep the states that satisfy φ". It can very well be the case that some of the states that satisfy φ may also satisfy its negation $-\varphi$. Clearly, this stems from the fact that negation $-$ in paraconsistent PAL is not classical, thus the methods of "eliminating the states that do not satisfy the announcement" and "keeping the states that satisfy the announcement" are not identical, unlike in classical logic. This distinction surfaces very clearly in paraconsistent PAL, and is one of the most important contributions of paraconsistent public announcement logic.

Let us now give a precise meaning to the public announcements. First, we define the updated model M' after the announcement the same way. Let $M = (S, \sigma, v)$ be a topological model where (S, σ) is a closed set topology where every $K \in \sigma$ is a closed set. For a formula $[\varphi]$, we obtain an *updated model* $M'_\varphi = (S', \sigma', v')$ where

$S' = S \cap |\varphi|$, $\sigma' = \{K \cap S' : K \in \sigma\}$, and $v' = v \cap S'$. We will remove the subscript when it is clear from the context.

Notice that there could also exist some other ways to revise the given model after an announcement. In other words, one may wish to exclude the states that satisfy the negation of the announcement from the space. We define $M^-_\varphi := (S \setminus |-\varphi|, \sigma', v')$ as the model obtained after the announcement of $[\varphi]$. We will call M^-_φ as the *reduced model*. Clearly, in classical logic, $M^-_\varphi = M'_\varphi$ for all models M and all formulas φ. But, in paraconsistent PAL, the reduced model is a subset of the updated model.

Lemma 9.1. *In classical PAL, for a model M, updated model M', and reduced model M^- are identical. In paraconsistent PAL, $M^- \subseteq M'$.*

Proof. Follows immediately from the definitions. □

Let us now present the formal aspects of paraconsistent public announcement logic, which we will call ParaPAL in short. We define the syntax of ParaPAL as follows for a propositional variable p and a falsum symbol $\bot$.

$$\bot \mid p \mid -\varphi \mid \varphi \wedge \varphi \mid \mathsf{K}\varphi \mid [\varphi]\varphi$$

As expected, K is the knowledge operator, and $[\varphi]$ denotes the public announcement of φ. We define disjunction and implication in the usual way. The dual operator L is defined as expected: $\mathsf{L}p := -\mathsf{K}-p$. For a more detailed exposition of multi-agent PAL in topological setting, see Başkent (2012). For simplicity, both in notation and exposition, we will only discuss the single-agent ParaPAL in this paper as extending it to a multi-agent case is straight-forward (Başkent 2013).

Let us give the semantics of ParaPAL now. Note that in ParaPAL, we have $|-p| = \mathsf{Clo}(S \setminus |p|)$. Also, $\bot$ is true nowhere (even if $p \wedge -p$ can be true). The semantics for propositional variables and Booleans are as usual. Let us reinstate the semantics of the modal and dynamic operators.

$$\begin{array}{lll} M, s \models \mathsf{K}\varphi & \text{iff} & \exists O \in \sigma.(s \in O \wedge \forall s' \in O : s', M \models \varphi) \\ M, s \models [\varphi]\psi & \text{iff} & M, s \models \varphi \text{ implies } M', s \models \psi \end{array}$$

In ParaPAL, the fact that after an announcement, the updated model will keep the states that satisfy the announcement and also may satisfy the *negation* of the announcement reflects the basic dictum of paraconsistent logic: Paraconsistent logics distinguish (at least) two different types of *true*s and *false*s. The trues that are only true and the trues that are also false; and similarly falses that are only false and the falses that are also true (Priest 1979). Therefore, in ParaPAL, after an announcement of φ, the agent comes to know φ (i.e. $M, s \models \mathsf{K}\varphi$), but we may also consider $-\varphi$ possible at the same state (i.e. $M, s \models \mathsf{L}-\varphi$).

An interesting observation here is that in ParaPAL, since the extension of *each* propositional variable is a closed set, we have $\mathsf{L}p \leftrightarrow p$. This observation follows from the topological fact that the closure of a closed set is already itself. That is, if the extension of each formula is a closed set already, its extension under the epistemic modal operator L will be the closure of the extension of the given

formula. But, the closure of a closed set is already itself by definition, therefore, the modal operator will not change the extension of a given formula yielding the logical equivalence $\mathsf{L}\varphi \leftrightarrow \varphi$ (Başkent 2013). Nevertheless, for expressivity purposes, we will keep the epistemic modal operator. This is a design decision similar to the classical PAL where the public announcement operator is not more expressive, yet provides succinctness (Kooi 2007). For convenience, we call the static fragment of ParaPAL (without the public announcement operator, but with the modal epistemic operator) as PTL after paraconsistent topological logic.

Before proceeding further, we need to make sure that the updated topology in ParaPAL is indeed a topology.

Lemma 9.2. *Given a closed set topology (S, σ). Then, for any p with a closed set extension, the updated space (S', σ') where $S' = S \cap |p|$ and $\sigma' = \{K \cap S' : K \in \sigma\}$ is also a topological space.*

Proof. The topology (S', σ') is indeed a well-known topology and called an induced topology. See Başkent (2012), for example, for a direct proof. □

The above lemma ensures that the semantics of public announcements in ParaPAL is well-defined.

9.3.2 Further Observations

9.3.2.1 Epistemic Modal Operator Is Redundant

An interesting result of PTL is that the epistemic operator is redundant. Nevertheless, for succinctness reasons, we keep the epistemic operator, as we already argued.

Lemma 9.3. *ParaPAL and PTL are equi-expressible.*

We will focus more on the reduction of ParaPAL to PTL in the next part.

Lemma 9.4. *ParaPAL is more expressive than PAL.*

In ParaPAL, we can have true statements such as $[p]\mathsf{K}(q \wedge -q)$. It would not be wrong to think that the introduction of impossible worlds to the model provides expressive richness for ParaPAL.

9.3.2.2 Reduction Axioms

Let us see whether the standard reduction axioms of classical PAL (which we gave in the Introduction) works in ParaPAL.

Consider the axiom $[\varphi]p \leftrightarrow (\varphi \rightarrow p)$ on a ParaPAL model $M = (S, \sigma, v)$ where $w \in S$, and p is a propositional variable. Suppose further that $M, w \models \varphi$.

$$\begin{array}{lll} M, w \models [\varphi]p & \text{iff} & M', w \models p \\ & \text{iff} & M, w \models p \\ & \text{iff} & M, w \models (\varphi \rightarrow p) \end{array}$$

Notice that the above result simply depends on the fact that the valuation of the propositional variables are independent from the topology.

ParaPAL presents a new negation. Thus, it is more important now to consider the reduction axiom for negation: $[\varphi]{-}\psi \leftrightarrow (\varphi \rightarrow -[\varphi]\psi)$. Similarly, take a ParaPAL model $M = (S, \sigma, v)$ where $w \in S$, and p is a propositional variable. Suppose further that $M, w \models \varphi$.

$$\begin{array}{lll} M, w \models [\varphi]{-}\psi & \text{iff} & M', w \models -\psi \\ & \text{iff} & w \in \mathsf{Clo}(S' \setminus |\psi|) \\ & \text{iff} & w \in \mathsf{Clo}((S \cap |\varphi|) \setminus |\psi|) \\ & & \text{as } w \in |\varphi| \text{ is assumed} \\ & \text{iff} & w \in \mathsf{Clo}(S \setminus (|\varphi| \cap |\psi|)) \\ & \text{iff} & w, M \models -[\varphi]\psi \end{array}$$

As we already pointed out, the reduction axioms for the epistemic modal operator holds vacuously. Thus, we obtain the following result.

Theorem 9.4. *ParaPAL reduces to PTL by the following reduction axioms:*

- $[\varphi]p \leftrightarrow (\varphi \rightarrow p)$
- $[\varphi]{-}\psi \leftrightarrow (\varphi \rightarrow -[\varphi]\psi)$
- $[\varphi]\psi \wedge \chi \leftrightarrow [\varphi]\psi \wedge [\varphi]\chi$
- $[\varphi]\mathsf{K}\psi \leftrightarrow (\varphi \rightarrow \mathsf{K}[\varphi]\psi)$

Proof. We already showed the soundness of the first two axioms. The third one on conjunction follows immediately, and the fourth one on the epistemic modality follows almost trivially as in ParaPAL and PTL the epistemic modality becomes redundant due to the properties of the closure operator (Başkent 2013). □

9.3.2.3 Topological Results

The most important advantage of adopting a topological background theory to express dynamic epistemic matters in a paraconsistent logic is to have the ability to make use of the topological properties of the model in understanding dynamic epistemic reasoning. In this section, we will consider various relevant topological concepts, and observe how they relate to expressing dynamic epistemologies.

Definition 9.3. A set X is called connected if $A \cap B \neq \emptyset$ whenever A, B are closed non-empty subsets and $X = A \cup B$. It is called totally disconnected if all of its subsets with more than one element are disconnected.

An interesting result for PTL models is the following.

Theorem 9.5 (Başkent 2012). *A PTL model with totally disconnected topology cannot be inconsistent.*

This theorem suggests a way to make the space consistent. It is also useful for our purposes in this paper. In other words, if the public announcement *disconnects* a space, then we can reduce the inconsistency to consistency by means of public announcements. The following theorem establishes the connection between inconsistent and consistent public announcement models via topological operations.

Theorem 9.6. *Let $M = (S, \sigma, v)$ be ParaPAL model where (S, σ) is an arbitrary topological space. Then if there exists a formula φ such that the topological space (S', σ') obtained after the announcement is totally disconnected, then $M'_\varphi = (S', \sigma', v')$ cannot be inconsistent.*

Proof. Given a ParaPAL model $M = (S, \sigma, v)$, call the updated model M'_φ. By reduction axioms, M'_φ reduces to a PTL model without changing the topology. Thus, if M'_φ, as a PTL model, is disconnected, by Theorem 9.5 it cannot be inconsistent. □

However, we should not over-read the above theorem. The existence of the public announcement φ that can turn arbitrary topological spaces to totally disconnected topological spaces is not guaranteed in each and every model.

A similar connection can be built between the static PTL and the dynamic ParaPAL.

Theorem 9.7 (Başkent 2012). *Let X be a connected topological space of closed sets with a PTL model on it. Then, the only subtheory that is not inconsistent is the empty theory.*

We can improve the above result within the context of ParaPAL as follows.

Theorem 9.8. *Let $M = (S, \sigma, v)$ be a ParaPAL model where (S, σ) is a connected topological space of closed sets. Then, the announcement of $\bot$ produces an updated model of M that has consistent theories.*

Proof. Let $M = (S, \sigma, v)$ be a ParaPAL model. We know that it is also a PTL model with the same topological structure. By Theorem 9.7, we know that the only theory that is consistent is the empty theory. In public announcement setting, we obtain this by announcing $\bot$ which is true nowhere. □

The above theorem is interesting. It reminds us that $\bot$ is nowhere true in paraconsistent spaces whereas some contradictions (in the form of $\varphi \wedge -\varphi$ for some φ) can be true somewhere. Additionally, it shows that the boundary points, the points that satisfy contradictions, are crucial to controls the inconsistencies. Concepts such as connectedness, as they relate to the boundary points, therefore play an essential role capturing inconsistent epistemologies in a dynamic setting.

An interesting aspect of topological PAL is whether/how the announcements stabilize the model, and how we can reach the limit models.

Definition 9.4. For a model M and a formula φ, define the announcement limit $\lim(M, \varphi)$ as the first model reached by successive announcements of φ that no longer changes after the announcement is made.

With static, ground Boolean formulas, the limit models are reached immediately after the first announcement. Moreover, in topological models for classical PAL, it is known that stabilization can take more than ω steps (Başkent 2012). This can also be seen as one of the strengths of topological models within the context of infinitary models. Then, the natural question is whether this property remains true in ParaPAL.

Theorem 9.9. *Model stabilization for ParaPAL models cannot take more than ω steps.*

Proof. The key point here is to observe that different definitions of common knowledge coincide in ParaPAL. This is usually the standard way to prove this statement (van Benthem and Sarenac 2004). As widely known, an announcement becomes a common knowledge after it is announced. Therefore a way to see how long the stabilization takes is to observe whether different definitions of common knowledge agree in ParaPAL.

Consider the following two definitions of common knowledge in Kripke models which we will only give in words, and refer the reader to van Benthem and Sarenac (2004) for a more detailed discussion.

- The reflexive and transitive closure of accessibility relations
- The fixed-point of the epistemic operator

In Kripkean models, these two definitions coincide as the knowledge modalities distribute over any arbitrary conjunctions. However, in PAL with classical topological semantics, these definitions do not coincide (van Benthem and Sarenac 2004; Başkent 2012).

On the other hand, in ParaPAL, since we have a closed set topology, and arbitrary intersections of closed set is still a closed set, we observe that the two definitions of common knowledge coincide, and they stabilize less than ω step. This can also be seen by the fact that the ParaPAL reduces to PTL losing is dynamic and epistemic modalities which make the stabilization *faster*. □

Another interesting direction is to observe how public announcements behave in some special inconsistent topological models. Now, we can turn into a well-known topological space, and observe how it affects the ParaPAL models. In Hausdorff spaces where distinct points have disjoint neighborhoods, we obtain the following results. Also note that, as a fact, in Hausdorff spaces compact sets are always closed.

Theorem 9.10. *Let $M = (S, \sigma, v)$ be a ParaPAL model where (S, σ) is a compact Hausdorff space. The stabilization for M takes less than ω steps.*

Proof. Let $M = (S, \sigma, v)$ be a ParaPAL model where (S, σ) is a compact Hausdorff space. Then, it is a closed set topology (thus, we do not need to impose it additionally). Since it is compact every arbitrary cover has a finite sub-cover. Thus, the stabilization, even if it takes more than ω-step can be converted into a stabilization with finitely many steps. □

Then, the next question is whether the PAL updates employ a continuous transformation in the model. Namely, given a ParaPAL model M and an arbitrary

formula φ, what is the connection between M and M'_φ in terms of continuous transformations? For this question, we use the functional representation of announcements, which we defined earlier. The following theorem holds immediately.

Theorem 9.11. *Every announcement is functionally representable in ParaPAL.*

Notice that, similar to the classical case of topological PAL, this result does not entail that f as above is truth preserving.

Now, we take one step further and consider the separation axiom $\mathbf{T}_6$ or perfectly normal spaces.

Definition 9.5 (Perfectly normal spaces). Given arbitrary closed sets K_1 and K_2 in a topology (S, σ). If there exists a continuous function $f : S \mapsto [0, 1]$ that separates K_1 and K_2 such that $f^{-1}(0) = K_1$ and $f^{-1}(1) = K_2$, then (S, σ) is called a perfectly normal topological space.

We then have the following theorem.

Theorem 9.12. *Let $M = (S, \sigma, v)$ be a ParaPAL model where (S, σ) is a perfectly normal topological space. If for two formulas φ and ψ, $M \not\models \varphi \wedge \psi$, then there exists a continuous transformation between M'_φ and M'_ψ.*

Proof. Let $M = (S, \sigma, v)$ be a ParaPAL model where (S, σ) is a perfectly normal topological space. Denote the extension of $|\varphi|^M = K_1$ and $|\psi|^M = K_2$. Then, as $M \not\models \varphi \wedge \psi$, we have $|\varphi \wedge \psi|^M = \emptyset$. Then, there exists a continuous function $f : S \mapsto [0, 1]$ that separates K_1 and K_2 such that $f^{-1}(0) = K_1$ and $f^{-1}(1) = K_2$ by definition.

Now, consider M'_φ and M'_ψ. In this case, observe that the carrier sets of M'_φ and M'_ψ are K_1 and K_2 respectively, again by definition. Thus, the transformation t from M'_φ to M'_ψ is given as follows:

$$t(x) = f^{-1}(f(x) + 1), \forall x \in K_1$$

The transformation t' from M'_ψ to M'_φ can also be defined similarly:

$$t'(y) = f^{-1}(f(y) - 1), \forall y \in K_2$$

By definition, t and t' are continuous. However, notice that, the transformation t is not truth preserving, nor a bisimulation. Therefore, the continuous transformation is, semantically, a renaming. □

Continuous transformation between two updated models mean that the topological (thus model theoretical) qualities of the two models are the same. Yet, since they may not have the same propositional valuation, these two models may not be bisimular.

In this section, we consider some topological concepts that are relevant to our discussion of paraconsistent public announcement logic. The field of topology is virtually unbounded, and it is possible to consider many other topological spaces and notions, and their impact on paraconsistent epistemologies.

9.4 Conclusion

Public announcement logic is an interesting playground to observe how epistemic reasoning based on paraconsistency works dynamically. Agents in ParaPAL can reason soundly in a world of inconsistencies. Our system is based on an inconsistent universe, yet takes announcements as honest and truthful epistemic operations.

The field is rich, and there can be considered a variety of future work possibilities including the algebraic connection between paraconsistency and public announcements, and paradoxical announcements. We leave it to future work.

Another interesting direction is the relation between mereology and public announcements. Mereology is the research area that studies the connection between parts and wholes, and exhibits intriguing algebraic qualities. Therefore, the question of how the relation between parts and wholes change after a public announcement is yet another interesting research direction to pursue.

References

Balbiani, P., Baltag, A., van Ditmarsch, H., Herzig, A., de Lima, T.: 'Knowable' as 'known after an announcement'. Rev. Symbol. Log. **1**(3), 305–334 (2008)

Başkent, C.: Public announcement logic in geometric frameworks. Fundam. Inf. **118**(3), 207–223 (2012)

Başkent, C.: Some topological properties of paraconsistent models. Synthese **190**(18), 4023–4040 (2013)

Başkent, C., Olde Loohuis, L., Parikh, R.: On knowledge and obligation. Episteme **9**(2), 171–188 (2012)

Carnielli, W.A., Coniglio, M.E., Marcos, J.: Logics of formal inconsistency. In: Gabbay, D., Guenthner, F. (eds.) Handbook of Philosophical Logic, vol. 14, pp. 15–107. Springer, Dordrecht/London (2007)

da Costa, N.C.A., Krause, D., Bueno, O.: Paraconsistent logics and paraconsistency. In: Jacquette, D. (ed.) Philosophy of Logic, vol. 5, pp. 655–781. Elsevier, Amsterdam (2007)

Gerbrandy, J.: Bisimulations on Planet Kripke. PhD thesis, Institute of Logic, Language and Computation, Universiteit van Amsterdam (1999)

Goodman, N.D.: The logic of contradiction. Z. für Math. Log. und Grundl. der Math. **27**(8–10), 119–126 (1981)

Kelly, T.: Disagreement, dogmatism and belief polarization. J. Philos. **105**(10), 611–633 (2008)

Kooi, B.: Expressivity and completeness for public update logics via reduction axioms. J. Appl. Non-class. Log. **17**(2), 231–253 (2007)

McKinsey, J.C.C., Tarski, A.: The algebra of topology. Ann. Math. **45**(1), 141–191 (1944)

McKinsey, J.C.C., Tarski, A.: On closed elements in closure algebras. Ann. Math. **47**(1), 122–162 (1946)

Mints, G.: A Short Introduction to Intuitionistic Logic. Kluwer, New York (2000)

Mortensen, C.: Topological seperation principles and logical theories. Synthese **125**(1–2), 169–178 (2000)

Plaza, J.A.: Logic of public communication. In: Emrich, M.L., Pfeifer, M.S., Hadzikadic, M., Ras, Z.W. (eds.) 4th International Symposium on Methodologies for Intelligent Systems, Charlotte, pp. 201–216 (1989)

Priest, G.: The logic of paradox. J. Philos. Log. **8**, 219–241 (1979)

Priest, G.: Paraconsistent belief revision. Theoria **67**(3), 214–228 (2001)
Priest, G.: Paraconsistent logic. In: Gabbay, D., Guenthner, F. (eds.) Handbook of Philosophical Logic, vol. 6, pp. 287–393. Kluwer, Dordrecht (2002)
Priest, G.: In Contradiction, 2nd edn. Oxford University Press, Oxford/New York (2006)
Priest, G.: An Introduction to Non-classical Logic. Cambridge University Press, Cambridge/New York (2008)
van Benthem, J., Bezhanishvili, G.: Modal logics of space. In: Aiello, M., Pratt-Hartman, I.E., van Benthem, J. (eds.) Handbook of Spatial Logics. Springer, Dordrecht (2007)
van Benthem, J., Sarenac, D.: The geometry of knowledge. In: Beziau, J.-Y., Facchini, A. (eds.) Aspects of Universal Logic. Volume 17 of Travaux Logic, pp. 1–31. Centre de recherches sémiologiques, Université de Neuchâtel, Neuchâtel (2004)

Chapter 10
Knowing Necessary Truths

Manuel Rebuschi

Abstract How account for the intuitive difference between simply knowing a necessary proposition, and knowing that it is a necessary truth? In the paper it will be shown that two-dimensional semantics does not do the job in an adequate way. A solution is provided which is based on Hintikka's worldlines. Assuming a slight extension of the syntax, modal epistemic logic can thus deal with classical puzzles like knowledge of identities.

Keywords Modal epistemology • Two-dimensional semantics • First-order modal logic • Hyperintensionality • Worldlines

10.1 Introduction

How account for the intuitive difference between simply knowing a necessary proposition φ, and knowing that φ is necessary? This is something modal epistemologists could be interested in. However, it seems that the most famous theoretical framework tailor-made for modal epistemology, i.e., *two-dimensional semantics* (2DS) *does not* account for such a difference. As far as we know, the issue was not even raised by 2DS-theoricians. In this paper, I propose a strategy to get hyperintensionality without such a disadvantage.

In Sect. 10.2, I will show that 2DS conflates knowledge of necessary propositions with knowledge that necessary propositions are necessary, and it will be argued that such a conflation *should* be avoided. In Sect. 10.3, I will thus present a reversal of perspective on the relationship between metaphysical and epistemic possibilities, so that the conflation *can* be avoided, and I will outline a enriched semantics that is more accurate to handle the two kinds modalities. I will conclude with a few remarks.

M. Rebuschi (✉)
L.H.S.P. – Henri-Poincaré Archives, University of Lorraine, Nancy, France
e-mail: manuel.rebuschi@univ-lorraine.fr

J. Redmond et al. (eds.), *Epistemology, Knowledge and the Impact of Interaction*,
Logic, Epistemology, and the Unity of Science 38, DOI 10.1007/978-3-319-26506-3_10

10.2 A Puzzle for Modal Epistemology

10.2.1 Blocking Modal Omniscience

It is well-known that standard epistemic logic (EL) assumes logical omniscience: since (EL) is an extension of propositional logic (PL), every PL valid formula φ is valid in EL too ($\vDash_{\mathsf{PL}} \varphi \Rightarrow \vDash_{\mathsf{EL}} \varphi$), and by *knowledge generalization* ($\vDash_{\mathsf{EL}} \varphi \Rightarrow \vDash_{\mathsf{EL}} K\varphi$, where $K\varphi$ stands for "*the agent knows that* φ"), every PL valid formula is known ($\vDash_{\mathsf{PL}} \varphi \Rightarrow \vDash_{\mathsf{EL}} K\varphi$).

It is also well-known that the situation of PL valid formulas (i.e., logically necessary truths, a.k.a. tautologies) generalizes to *necessary truths*, whether metaphysical or analytical. Indeed, put in the framework of possible-world semantics, a proposition φ is known by an agent as long as it is true in all her *epistemically possible worlds*, i.e., in all the possible worlds epistemically indistinguishable from the agent's perspective. Hence necessary truths, which are true in every possible world *period*, are expected to be true in every epistemically possible world. So using a bimodal logical language $\mathcal{L}(\Box, K)$, the following generally holds:

($\blacklozenge$) $\Box\varphi \Rightarrow K\varphi$ *Modal omniscience*

(where $\Box\varphi$ stands for "*it is necessary that* φ".)

Several strategies were developed to avoid such an odd consequence, like *impossible-world semantics* (IWS) (Rantala 1982) and 2DS (Chalmers 2004). IWS is a nice way to avoid logical omniscience by the admission of logically impossible worlds that are nonetheless epistemically possible according to an agent: the agent conceives as possible, situations where some logical consequences of her knowledge do not obtain, hence situations which are impossible.[1]

2DS proceeds differently. Roughly said, the idea is to evaluate sentences relatively to a pair of worlds, i.e., to add a world considered as actual to the current one. However, even though 2DS sticks to a fixed set of metaphysically possible worlds, the resulting situation is analogous to that of IWS: some worlds are (epistemically) considered as actual while not being so, giving rise to an inflation of possible situations far beyond the initial set of possible worlds.

[1]Technically, standard Kripkean models are extended in order to include worlds where the valuation is no more standard. In such worlds, the value of complex formulas is no more calculated according to that of atoms, but it is just postulated. For instance, the conjunction of p, $p \rightarrow q$ and $\neg q$ can obtain at the same world. If such a world is epistemically accessible from some (standard) world w, then at w the agent can be said to know that p and that $p \rightarrow q$, while ignoring that q. This makes such non-standard worlds look impossible, even though they are consistent theoretic constructions.

Table 10.1 2DS evaluation of "Water = H_2O"

	w_0	w_1	w_2	...
w_0* ("water" refers to H_2O)	T	T	T	...
w_1* ("water" refers to H_2O)	T	T	T	...
w_2* ("water" refers to XYZ)	F	F	F	...
...	...	...	...	...

10.2.2 2DS Threats on Metaphysics

According to both strategies, one can then consider a proposition φ which is necessary ($\Box\varphi$) but unknown ($\neg K\varphi$). Now it seems that even in a framework like 2DS, if a proposition ψ is necessarily true ($\Box\psi$) and known by an agent ($K\psi$), then the agent ipso facto knows that it is necessarily true ($K\Box\psi$). So even if ($\blacklozenge$) is blocked, a new inference:

($\star$) $\Box\psi, K\psi \Rightarrow K\Box\psi$ *Metaphysician's omniscience*

is not. This can be shown quickly considering the 2DS-style Table 10.1.

Here, w_0 is the actual world. The columns are composed of w_0 and of counterfactual worlds w_1, w_2, ..., i.e., of metaphysically possible worlds. These are the usual possible worlds where sentences are evaluated. The different rows form the second dimension of 2DS: they are made of the previous possible worlds *considered as actual*, relatively to which the extension of "water" is determined, w_0*, w_1*, w_2*...

Following Stalnaker's (1978) metasemantic interpretation, one can see the worlds-considered-as-actual as contexts of utterance. In w_0* and w_1*, the utterance of "water" is meant to designate water, but in w_2* it is meant to designate another watery stuff, XYZ – like on Putnam's Twin-Earth (Putnam 1975). After Chalmers (2004), I will consider the worlds-considered-as-actual as the epistemic possibilities of an agent. As far as the agent does not know which stuff is designated by "water", she will consider w_2* as a genuine epistemic possibility.[2]

Since "water" is a natural-kind term, it rigidly refers to its extension. If its extension is water, as is the case as uttered in the actual world w_0* or in w_1*, then the sentence "Water = H_2O" is true in every counterfactual possibility – whereas if the same expression is used to designate the watery stuff XYZ, like in w_2*, the sentence will be false in any counterfactual possibility.

Metaphysical necessities like $\Box$(water = H_2O) are evaluated along each horizontal line: on the first line, "water = H_2O" is true in every counterfactual possibility (on the same line), so $\Box$(water = H_2O) is true if w_0 is considered as actual, i.e. at w_0* – and similarly for w_1*; on the third line, the formula is false (since "water" is used to designate XYZ). By contrast, epistemic statements like K(water = H_2O), are evaluated along the diagonal, i.e., among the set of world-pairs

[2] A more complete overview of 2D semantics and its various interpretations is provided by the SEP entry (Schroeter 2012).

$\{\langle w_0{}^*, w_0\rangle, \langle w_1{}^*, w_1\rangle, \langle w_2{}^*, w_2\rangle, \ldots\}$. To put it in words: they are evaluated at every world considered as actual. In the above example, the sentence "Water = H_2O" is not true at $\langle w_2{}^*, w_2\rangle$, so the agent cannot be said to know that it is true: the statement K(water = H_2O) is false.

What happens if the agent learns that water is made of H_2O? On this 2DS modeling, we will get rid of all the epistemic possibilities incompatible with the statement, like the situation where w_2 is considered as actual. In the above table, the third column which correspond to a metaphysical possibility no more compatible with what the agent knows would be eliminated.[3] Then the third line would also be deleted. The desired upshot is that on the diagonal, "Water = H_2O" is always true, which gives the truth-conditions of K(Water = H_2O). Another immediate but unwanted consequence of this cleaning is that in every epistemic possibility, the identity "Water = H_2O" is now necessarily true, i.e., $\Box$(water = H_2O) holds. But this implies that the agent knows it: $K\Box$(water = H_2O).

10.2.3 Modal Knowledge for Free?

It is worth noticing that Kripke himself seems a bit confusing about knowledge of necessary truths:

> All the cases of the necessary *a posteriori* advocated in the text have the special character attributed to mathematical statements: philosophical analysis tells us that they cannot be contingently true, so any empirical knowledge of their truth is automatically empirical knowledge that they are necessary. (Kripke 1972)

In this passage, Kripke apparently does not make any substantial distinction between knowing a necessary truth, and knowing that this truth is necessary. Kripke would thus endorse the Metaphysician's omniscience ($\star$), assuming an "automatic" transition from the former to the latter. Nevertheless, the quotation seems self-contradictory because the transition is also said to be granted by "philosophical analysis". So what? Is empirical knowledge sufficient or not to gain knowledge that a proposition is necessary?[4]

A sufficient condition for an automatic transition could be that an agent knows the boundaries of the set of metaphysical possibilities. Then a necessary truth ψ could be considered as necessary as soon as it would be known, since the agent

[3] This could be done by an updating of an accessibility relation between possible worlds. However, I will not go into details about the formal implementation of the general idea.

[4] According to a more charitable reading, Kripke would mean that once philosophical analysis is done e.g. for natural kind terms in general, every empirical knowledge of, say, identities (like the identity between water and H_2O) would automatically lead to a knowledge that these identities are necessary. But it means that the analysis must have been done by the knowing agent herself, and it remains odd to qualify her knowledge about the modal status of a truth as *empirical* – whereas knowledge of necessary identities can of course be empirical.

could explore the possibilities beyond her epistemic possibilities (i.e., she could explore those situations that are incompatible with what she knows, but nonetheless that could have been the case), and check whether ψ would hold there. But if such knowledge were required, it seems that all the benefits of 2DS fade away! Indeed, knowing what the boundaries of the sphere of metaphysical possibilities are would entail knowing that any necessary truth is necessary; it would also entail knowing any necessary truth, which is modal omniscience ($\blacklozenge$).

Hence knowledge of what is metaphysically possible is sufficient, but certainly not necessary. Whatever be the adequate explanation, the problem raised by the inference ($\star$) should appear in an obvious manner: agents should not get its conclusion for free. Knowing the modal status of a truth is not a trivial affair, it requires metaphysical reasoning (hence the label of *Metaphysician's omniscience*). Many cases can be considered where the transition is not automatic at all. Children learning arithmetic can be said to know many necessary truths, like: $2 + 3 = 5$, but surely not that they are necessary. Other examples are of course provided by empirically known identities, which were not considered necessary before Kripke's account. So the inference ($\star$) must be either blocked or circumvented.

10.2.4 Which Diagnosis for 2DS?

2DS is unable to discriminate between epistemic possibilities and conceptual possibilities. However what is known to be false, and consequently eliminated from the epistemic possibilities, can be considered either conceptually possible or not. That such a distinction is not carefully taken into consideration by the two-dimensional framework (Fiocco 2007), is maybe one of the reasons why the Metaphysician's omniscience obtains.

But there is a more crucial reason. To a certain extent, the strategy of 2DS regarding metaphysical necessity is similar to that of IWS: taking into account situations that are epistemically (or subjectively) conceivable, although not objectively possible. Even though there is strictly speaking no addition of possible worlds, epistemic possibilities are handled via an expansion of the set of situations. It is basically an inflationist account regarding possibilities.

This theoretical shaping has an obvious consequence. When knowledge is updated by the acquisition of new information φ, some of the previous epistemic possibilities are deleted: precisely those which are incompatible with φ. Now the decreasing of epistemic possibilities would ultimately lead us back to the original situation, i.e., to the original set of metaphysically possible worlds, hence to the Metaphysician's omniscience.

Is it the price to pay for hyperintensionality? Another strategy would require incomplete worlds for the agent. To go back to the above example, a situation like that described in Table 10.2 would circumvent the Metaphysician's omniscience. But this is not allowed in the 2DS framework: the values of the expressions in every possible world are independent from the idiosyncrasies of any agent.

Table 10.2 2DS with indefinite worlds

	w_0	w_1	w_2	...
w_0* ("water" refers to H_2O)	T	T	T	...
w_1* ("water" refers to H_2O)	T	T	**?**	...
...	...	...	...	...

10.3 Reversing the Perspective

10.3.1 An Alternative Strategy

The proposal of this section does not do the job to block logical omniscience, but it is enough to avoid modal omniscience ($\blacklozenge$). Rather than expanding the space of possibilities, we shall stick to our fixed set of metaphysically (objective) possible worlds. The set of epistemically possible worlds is thus considered as a proper part of that of metaphysically possible worlds. The fine-grained description of an epistemic agent is accounted for using Hintikka's *worldlines* (Hintikka 1967, 1969). As a result, we have a relational structure of possible worlds augmented with a system of relations between the individual objects.

Worldlines formally correspond to Montague's individual concepts. According to Hintikka nevertheless, it is a short-eyed point of view to reduce worldlines to individual concepts, since the latter are purely linguistic whereas worldlines are independent from language. Such worldlines are understood as constituting individuals in our conceptual scheme in Hintikka's conception. However, I won't stick to Hintikka's specific metaphysics of transworld individuals – my objective is only to use the formalism that can result from such an addition. However, though not being an orthodox follower, I retain from Hintikka's conception the idea that worldlines encode the agents' ways of identifying objects through possible worlds. This is basically a non-linguistic issue, but it immediately interacts with language as one considers attitude ascriptions. This is indeed our point of departure.

I can state that Michelle Obama knows that Barack Obama is president, and that Hollande knows it too. Of course, the ways of identification of Barack Obama need not coincide between Michelle Obama, Hollande, and myself. Even though the name "Barack Obama" can be considered a rigid designator, i.e. picking out one and the same individual at every metaphysically possible world, this is not the case when used to represent the ways Michelle Obama, Hollande, or myself, identify Barack Obama in different possible worlds. For then, it will vary and depend of our respective knowledge of Barack Obama, be it partly direct (by acquaintance) or purely descriptive. So for a sentence like "Michelle Obama knows that Barack Obama is president" uttered by myself, the name can be interpreted according either to the ascriber's worldline, to the ascribee's.[5]

[5]Such ascriptions do not presuppose any language use by the ascribee. One could also state the following about the Obama family's pet dog: "Bo knows that Barack Obama is in the kitchen".

My proposal will thus involve two kinds of values for names (i.e. for individual constants). First the interpretation function will carry a fixed value – the same individual object for every possible world – so that names behave like rigid designators. Second, we will convey worldlines as values so that names behave like flexible designators, in order to account for their use in attitude ascription contexts – and ultimately, to account for the agents' ways of identification of individuals. Let us go into formal details.

10.3.2 An Alternative Semantics

I will now sketch a new semantics for first-order modal logic, in the line of Kraut (1983), Gerbrandy (2000), Aloni (2005) or Tulenheimo (2009).

Definition 10.1 (Syntax). Terms and formulas of the first-order bimodal language $\mathcal{L}(\Box, K)$ are defined as follows:

$$\begin{array}{ll} \text{Terms:} & t ::= a \mid x \\ \text{Formulas:} & \varphi ::= \top \mid Rt_1 \ldots t_n \mid \exists x\, \varphi \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid K\varphi \mid \Box\varphi \end{array}$$

where a is an individual constant, x an individual variable, and R a n-ary relation symbol.

Definition 10.2. An *enriched Kripke model* for a first-order bimodal language $\mathcal{L}(\Box, K)$ is a tuple $\mathbf{M} = \langle W, @, R_\Box, R_K, D, I, I_K \rangle$, where:

- W is a non-empty set of possible worlds;
- $@$ is a distinguished world ("*the actual world*");
- $R_\Box \subseteq W \times W$ is the accessibility relation between metaphysically possible worlds;
- $R_K \subseteq R_\Box$ is the accessibility relation between epistemically possible worlds;
- D is a domain of individuals[6];
- I is an interpretation function that assigns an individual $I(c) \in D$ to each individual constant c, and a subset $I(P, w)$ of D^n to each n-ary predicate P and possible world w;
- I_K is a function that maps every individual constant c onto a *worldline* $I_K(c)$, which is a (possibly partial) function from possible worlds to individuals ($I_K(c) : w \longmapsto I_K(c)(w) \in D$, if defined on w), such that $I_K(c)(@) = I(c)$.

This statement could involve either the ascriber's way of identifying the bearer of the name, or the ascribee's, which is by no way linguistic.

[6]For simplification the domain is supposed to be constant across possible worlds; to get variable domains, D should be defined as a function ascribing a domain of individuals D_w to each possible world $w \in W$.

Everything is standard in enriched models, except the consideration of the I_K function that maps individual constants onto worldlines. Such a function extensionaly coincides with the standard interpretation function at least in the actual world @ – i.e., in the actual world the worldline corresponding to a given individual constant picks out the very object referred to by that constant.

The inclusion $R_K \subseteq R_\Box$ means that no world is added to the set of metaphysically possible worlds. Hence we account for modal errors as misconceivability, rather than as conceivability of possibilities beyond metaphysical possibilities. We thus are in accordance with the claim, suggested by Hume, that nothing (genuinely) conceivable is absolutely impossible. The inclusion of the two relations entails constraints on their respective properties. For instance one can consider $R_\Box$ as a total relation (hence respecting S5) and R_K as only reflexive and transitive (hence respecting S4), but the reciprocal would not be allowed.

10.3.3 Example

One can consider the case pictured in Fig. 10.1. Three metaphysically possible worlds are represented, u, v and w, the first two of which also being epistemically possible. For simplicity the accessibility relations $R_\Box$ and R_K are not drawn, but we can assume that they are equivalence relations and relate u, v and w on the one hand, and u and v on the other.

According to the model, it is metaphysically necessary that Pa, because $I(a)$ belongs to the extension of P in every metaphysically possible world u, v and w;

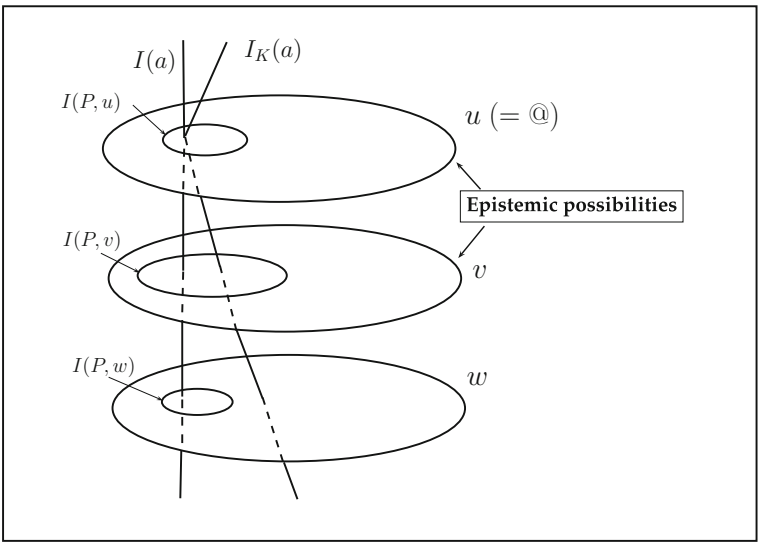


Fig. 10.1 A model without Metaphysician's omniscience

the agent knows that Pa, because the value of $I_K(a)$ belongs to the extension of P in every epistemically possible world u and v; however, the agent does not know that it is necessary that Pa since the value of the worldline $I_K(a)$ in w does not belong to the local extension of P.

10.3.4 Extending the Language

The values of the individual constants are expected to vary according to the embedding modal operator: the value of a is $I(a)$ if the constant is at most in the scope of $\Box$, whereas it must be $I_K(a)$ when occurring in the scope of K. Defining the truth-conditions of our formulas is thus rather uneasy. In order to get a semantics that remains compositional, we extend the language and let two different symbols occur instead of a, one for each semantic value. More accurately:

– we will put aside a for the rigid designators: KPa will mean "The agent knows that a is a P", where "a" corresponds to the way the ascriber denotes the bearer of the name;
– we will use a^K for possibly flexible designators: KPa^K will mean "The agent knows that a is a P", where "a" corresponds to the way the agent identifies the bearer of the name.

It is questionable to allow for different terms for a single individual constant, corresponding to a single proper name in natural language. Are we thus committed to postulating a general ambiguity in natural language? There is a sense in which a similar ambiguity is widely acknowledged: the ambiguity between *de dicto* and *de re* attitudes, which is also accounted for within the present framework (see below).

However, it cannot be the whole answer. What is taken into account here is the possibility in principle to label every individual constant a with differences from agent to agent: a^{K_i} for agent i, a^{K_j} for agent j, etc. For instance, if an agent i knows that another agent j knows that Pa, this could be done using two formulas: either $K_iK_jPa^{K_i}$, or $K_iK_jPa^{K_j}$. It seems unrealistic to consider that at the level of "logical forms" there should be as many names as there are speakers.

Of course we should consider that there are (at least) as many worldlines as agents, but we would like to avoid a direct translation of such a fact into the syntax. Actually, the job can be done using other notations, like independent quantifiers of Hintikka's IF epistemic logic (Hintikka 2003) or subjunctive markers of Wehmeier's conception (Wehmeier 2004).[7] Nevertheless, since the two other notations presuppose alternative general theories of modalities, I will use the unorthodox one just sketched above.

[7] KPa as it is used in the present paper is equivalent to the IF formula $KP(a_{/K})$, and would be unchanged in Wehmeier's notation; KPa^K would correspond to KPa and to KPa^s respectively.

Definition 10.3 (Extended syntax). Terms and formulas of the extended language $\mathcal{L}'(\Box, K)$ are defined as follows:

$$\begin{array}{ll} \text{Terms:} & t ::= a \mid a^K \mid x \\ \text{Formulas:} & \varphi ::= \top \mid Rt_1 \ldots t_n \mid \exists x\, \varphi \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid K\varphi \mid \Box\varphi \end{array}$$

where a is an individual constant, x an individual variable, and R a n-ary relation symbol.

10.3.5 Worldlines

We will consider the values $I(c)$ as *rigid total worldlines*, i.e., as constant functions from W to D; the notion of rigid total worldline is then extended to every individual in the domain D (i.e., to the individuals that are not denoted by any individual constant). Let us denote by $\overline{D}$ the set of rigid total worldlines so generated: $\overline{D} = \{f \in D^W \mid \forall w \forall w'\, f(w) = f(w')\}$.

In addition, we consider *(possibly) flexible worldlines* that correspond to the agent's ways of transworld identification. They constitute a subset of all the (partial and total) functions from W to D: $\mathsf{FWL} \subset \{f \in D^X \mid X \subseteq W\}$. The set of worldlines WL is thus the union of all the rigid total worldlines generated from the domain with the chosen set of possibly flexible worldlines (ultimately corresponding to ways of identification): $\mathsf{WL} = \overline{D} \cup \mathsf{FWL}$.

We define *assignment functions* g as functions mapping the variables onto WL.

Accordingly, the usual quantifiers $(\forall, \exists)$ will also take their values in the set of worldlines WL. Terms and formulas are interpreted relatively to an enriched Kripke model $\mathbf{M}$, a possible world w, and an assignment function $g : Var \to \mathsf{WL}$.[8]

Definition 10.4. Value of terms:

$$\begin{array}{lcl} [x]_{\mathbf{M},w,g} & = & g(x)(w), \text{ where } x \text{ is a variable;} \\ [a]_{\mathbf{M},w,g} & = & I(a), \text{ where } a \text{ is an individual constant;} \\ [a^K]_{\mathbf{M},w,g} & = & I_K(a)(w), \text{ where } a \text{ is an individual constant.} \end{array}$$

Definition 10.5. Interpretation of formulas:

$$\begin{array}{lll} \mathbf{M}, w, g \vDash Pt_1 \ldots t_n & \text{iff} & \langle [t_1]_{\mathbf{M},w,g}, \ldots, [t_n]_{\mathbf{M},w,g} \rangle \in I(P, w) \\ \mathbf{M}, w, g \vDash t_1 = t_2 & \text{iff} & [t_1]_{\mathbf{M},w,g} = [t_2]_{\mathbf{M},w,g} \\ \mathbf{M}, w, g \vDash \neg\varphi & \text{iff} & \mathbf{M}, w, g \nvDash \varphi \\ \mathbf{M}, w, g \vDash \varphi \wedge \psi & \text{iff} & \mathbf{M}, w, g \vDash \varphi \text{ and } \mathbf{M}, w, g \vDash \psi \\ \mathbf{M}, w, g \vDash \exists x\, \varphi & \text{iff} & \exists \ell \in \mathsf{WL} \text{ such that: } \mathbf{M}, w, g[x/\ell] \vDash \varphi \\ \mathbf{M}, w, g \vDash \Box\varphi & \text{iff} & \text{for all } w', \text{ if } wR_{\Box}w' \text{ then } \mathbf{M}, w, g \vDash \varphi \\ \mathbf{M}, w, g \vDash K\varphi & \text{iff} & \text{for all } w', \text{ if } wR_K w' \text{ then } \mathbf{M}, w, g \vDash \varphi. \end{array}$$

[8] In what follows, I use the notations of Aloni (2005).

Remark. Since worldlines need not be total functions it can happen that an atomic formula is evaluated at a world where the value of one of its terms is not defined. For example, Pa^K can be evaluated at $\langle \mathbf{M}, w, g \rangle$ whereas $I_K(a)$ is undefined at w (and similarly for Px and $g(x)$). According to our definition the formula is not satisfied, so its negation is, and this situation leads to well-known difficulties. The definition would thus require refinements. One possibility is to extend each partial worldline so that when undefined it picks out always the same single abstract object added to the domain D; if the interpretation of predicates were extended accordingly, one could then avoid truth-value gaps. Then one's intuitions about negation must be accordingly refined.[9]

10.3.6 Application

Let us go back to the example of Fig. 10.1. According to our definition, we have:

$\mathbf{M}, u, g \models \Box Pa$

because in every metaphysically possible world, i.e. in u, v and w, the value of a, i.e. $I(a)$ belongs to the local extension of P, resp. $I(P, u)$, $I(P, v)$ and $I(P, w)$. We also have:

$\mathbf{M}, u, g \models KPa$

because the value of a belongs to the local extension of P in every epistemically possible world, i.e. to $I(P, u)$ and $I(P, v)$. But this is not so interesting (this is true of Pa and of any necessary formula). We had rather consider:

$\mathbf{M}, u, g \models KPa^K$

that also obtains: in u and v (only), the local values of $I_K(a)$ (i.e., $I_K(a)(u)$ and $I_K(a)(v)$) belong to the local extensions of P. However, it could have been the case that the formula were *not* satisfied, just if the local value of $I_K(a)$ in some epistemically possible world did not belong to the local extension of P – it means that like 2DS, our account can avoid modal omniscience.

Now the agent does not know that Pa necessarily holds:

$\mathbf{M}, u, g \not\models K\Box Pa^K$

since there is a possible world, v, such that uR_Kv, and there is another possible world, w, such that $vR_\Box w$ and $I_K(a)(w) \notin I(P, w)$. Of course, the formula would be satisfied if the agent had another worldline for a, $I'_K(a)$, whose local value at w would belong to $I(P, w)$. This further requirement corresponds to the supplementary information that must be possessed by the agent: information about the metaphysically possible properties of the individual (she believes to be) a.

[9] I thank a referee for having raised this issue in my original definition.

We can also consider the case of quantified formulas (this is the kind of basic case accounted for in Aloni 2005):

$\mathbf{M}, u, g \models \exists x\,(x = a \wedge KPx)$

The formula holds: there is a worldline in WL, namely $I_K(a)$, such that (i) its value in u ($I_K(a)(u)$) coincides with that of a ($I(a)(u) = I(a)$), and (ii) in every epistemically possible worlds, i.e., in u and v, its local value is in the local extension of P. With quantified formulas one can thus use worldlines without proper names, as ways of identifying individuals independently from language, hence in accordance with Hintikka's perspective.

The semantics of worldlines offers other benefits. As accounted for in Kraut (1983) it can be used to contrast *de dicto* and *de re* knowledge: in order to get *de re* knowledge, the further requirement is that the worldline be *constant* through the accessible epistemically possible worlds.[10] However, such worldlines *do not* collapse onto rigid designators: they are not required to be constant, and even not to be defined, on the metaphysically possible worlds which are situated beyond the epistemic possibilities.

10.3.7 Back to Omniscience

What about the initial inferences ($\blacklozenge$) and ($\star$)? Both of them can hold! It immediatly follows from the definitions, especially from the inclusion $R_K \subseteq R_\Box$, that if φ is a necessary truth then it is known: $\Box\varphi \Rightarrow K\varphi$ ($\blacklozenge$). And assuming that the accessibility relation $R_\Box$ is transitive, it can easily be checked that when a truth is necessary, $\Box\varphi$, then it is necessarily the case, i.e. $\Box\Box\varphi$ (thanks to transitivity), hence it is known that it is necessary, $K\Box\varphi$; so ($\star$) obtains.

Fortunately, in the aforementioned example both inferences appear to be innocent. Indeed, KPa does not properly formalize the knowledge of Pa in general, but only *de re* knowledge: the knowledge of a that it is a P. Let us consider for instance the case of Hesperus (h) and Phosphorus (p). The formula: $K(h = p)$ means that the agent knows (*de re*) of Hesperus and of Phosophorus that they are identical; to put it in other (equivalent) words, the formula means that the agent knows of Venus that it is identical with itself. This is trivially true. In a sense, such a knowledge can follow from $\Box(h = p)$ (according to ($\blacklozenge$)), and the two formulas can entail $K\Box(h = p)$ (according to ($\star$)), i.e., the knowledge that the identity of Venus with itself is necessary.

[10] So this is not a matter of scope, and the formula $\exists x\,(x = a \wedge KPx)$ can be used to ascribe *de dicto* knowledge – as far as the worldline picked out by the existential quantifier is variable.

This must be contrasted with $K(h^K = p^K)$, which means that the agent has two worldlines (i.e. two ways of identifying Venus), one for each name, which provide the same value in every epistemically possible world. *This* corresponds to the genuine knowledge of the identity between Hesperus and Phosphorus. It neither is implied by $\Box(h = p)$,[11] nor implies that $K\Box(h^K = p^K)$.[12]

10.4 Conclusion

Shall we give up? Could we get rid of the rigid mapping of individual constants, and define everything in terms of worldlines? This option is generally favored by Hintikka according to whom Kripke's rigid designators are meaningless. But what would be individuated if nothing were beforehand given? To that respect Hintikka's conception leads to antirealism, unless one endorses his metaphysics of worldlines as genuine individuals rather than as mere conceptual means (i.e. ways of identification of objects). However, it appears that postulating rigid designators is indispensable to determine transworld identity between individuals when no individuating criterion is at disposal, as is the case for alethic possibility (Rebuschi 2009).

The framework I proposed in this paper combines metaphysical possibilities à la Kripke with epistemically possible worlds à la Hintikka. In a way, it is a two-dimensional framework. However, it provides several advantages in comparison with 2DS. One of them is the way it enables us to circumvent the inference ($\star$) of perfect metaphysicians considered in this paper. Hence there is still some work to be done by metaphysicians to distinguish between necessary and contingent propositions in our knowledge.

Acknowledgements Preliminary versions of this paper were presented at conferences in Rennes (France) and Rijeka (Croatia). I wish to thank Filipe Drapeau-Contim, Ghislain Guigon, Pierre Joray, Pascal Ludwig, Claudine Tiercelin, Tero Tulenheimo, and an anonymous reviewer for helpful comments and suggestions.

[11] It *is* implied by $\Box(h^K = p^K)$, but in general this formula is not true, since the two names are expected to encode two diverging worldlines – two different senses, to put it in Fregean terms.

[12] A complete theory would allow worldlines for predicates and not only for names. See Egré (2014) for a proposal in that direction. It would expand my solution to cases with no individual constants, like the ascription of knowledge of "Tigers are mammals". That this is a necessary truth (after Putnam) does not mean that knowing this truth implies knowing that it is necessary. However, a referee stressed that the inference ($\star$) would not be blocked in the propositional case. This is true, and it shows that propositional modal logic is not fine-grained enough to handle the distinction between knowledge of a necessary truth and knowledge that this truth is necessary.

References

Aloni, M.: Individual concepts in modal predicate logic. J. Philos. Log. **34**, 1–64 (2005)
Chalmers, D.J.: Epistemic two-dimensional semantics. Philos. Stud. **118**, 153–226 (2004)
Egré, P.: Hyperintensionality and de re beliefs. In: Lihoreau, F., Rebuschi, M. (eds.) Epistemology, Context, and Formalism. Synthese Library, vol. 369, pp. 213–243. Springer, Dordrecht (2014)
Fiocco, M.O.: Conceivability and epistemic possibility. Erkenntnis **67**, 387–399 (2007)
Gerbrandy, J.: Identity in epistemic semantics. In: Cavendon, L., et al. (eds.) Logic, Language and Computation, vol. 3, pp. 147–159. CSLI, Stanford (2000)
Hintikka, J.: Individuals, possible worlds, and epistemic logic. Noûs **1**, 33–62 (1967)
Hintikka, J.: Semantics for propositional attitudes. In: Davis, J.W., Hockney, D.J., Wilson, W.K. (eds.) Philosophical Logic, pp. 21–45. D. Reidel, Dordrecht (1969)
Hintikka, J.: A second generation epistemic logic and its general significance. In: Hendricks, V.F., et al. (eds.) Knowledge Contributors, pp. 33–55. Kluwer, Dordrecht (2003)
Kraut, R.: There are no de dicto attitudes. Synthese **54**, 275–294 (1983)
Kripke, S.: Naming and Necessity. Harvard University Press, Cambridge (1972)
Putnam, H.: The meaning of 'meaning'. In: Gunderson, K. (ed.) Language, Mind and Knowledge. University of Minnesota Press, Minneapolis (1975)
Rantala, V.: Impossible worlds semantics and logical omniscience. Acta Philosophica Fennica **35**, 106–115 (1982)
Rebuschi, M.: Modalités épistémiques et modalités aléthiques chez Hintikka. Revue Internationale de Philosophie **250**, 395–404 (2009)
Schroeter, L.: Two-dimensional semantics. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy (Winter 2012 edn.) (2012). http://plato.stanford.edu/archives/win2012/entries/two-dimensional-semantics/
Stalnaker, R.: Assertion. Syntax Semant. **9**, 315–332 (1978)
Tulenheimo, T.: Remarks on individuals in modal contexts. Revue Internationale de Philosophie **250**, 383–394 (2009)
Wehmeier, K.F.: In the mood. J. Philos. Log. **33**, 607–630 (2004)

Chapter 11
Modified Tableaux for Some Kinds of Multimodal Logics

Emilio Gómez-Caminero and Ángel Nepomuceno

Abstract A multimodal logic is a logic where a certain number of different modal operators appear. In some of these logics we can have at our disposal a labeled tableaux method whereby different modal operators give rise to different labels. The properties of the accessibility relations, in the semantic view, may be treated by means of what we call *inheritance rules*.

The easiest cases are those in which all modal operators are of the same type, such as multiagent epistemic or doxastic logic. In these cases we can propose a modular tableau method that we can adapt to the most important systems only changing the inheritance rules. Although some of these systems give rise to infinite branches, we can avoid the infinity by means of some restrictions in the use of rules. More complicated cases require additional rules to deal with the relationship between different modal operators. Finally, some infinitary operators, such as *common knowledge* or *sometime*, may be dealt with using DB-tableaux or recursive rules.

Keywords Multimodal logics • Tableaux methods • Labeled tableaux • Inheritance rules • DB-tableaux • Recursive rules

11.1 Introduction

According to the fundamental property of classical semantic tableaux, a (finite) set of formulas is satisfiable if and only if the tableaux whose root is this set is open. As a corollary, any formula is valid if and only if the tableaux whose root is the negation of that formula is closed. In decidable logics the method is very suitable, though sometimes, as in classical first order logic, the standard procedure gives rise to infinite branches, which has been dealt with by the modification of certain rules.

When the method is applied to study other logics, some problems arise. So, in modal logics, to construct tableaux, the corresponding accessibility relation has to be taken into account, that is to say, formulas must be labeled with the names of

E. Gómez-Caminero (✉) • Á. Nepomuceno
Grupo de Lógica Lenguaje e Información, Universidad de Sevilla, Seville, Spain

J. Redmond et al. (eds.), *Epistemology, Knowledge and the Impact of Interaction*,
Logic, Epistemology, and the Unity of Science 38, DOI 10.1007/978-3-319-26506-3_11

worlds according to the features of those relations. In this paper, we shall study these problems as they relate to multimodal logics.

This paper is organized as follows. After this introduction, in the Sect. 11.2 we tackle the basic case, in which the only operators considered in the language are the standard ones, namely $\Box$ and $\Diamond$. Section 11.3 is devoted to presenting the two kinds of rules, which have been named as *common* and *inheritance rules*, with respect to the most well-known multi-agent systems. Then in the fourth section we introduce a combination of different kinds of modalities in order to tackle doxastic-epistemic systems. The fifth section is devoted to the study of knowledge in groups of agents and defining rules for operators of "knowledge of everybody" and "common knowledge". In the sixth section, the objective is to explain the tableaux method for temporal logic. To finish, the last section is devoted to concluding remarks in which future lines of work are pointed out.

11.2 The Basic Case

The basic case is the one in which we have only one modal operator $\Box$ and its dual $\Diamond$. These operators may be interpreted as alethic operators, deontic operators, etc., giving rise to different systems of modal logic. In the semantic view, the set of valid formulas for each of these logics depends on the properties of the accessibility relation.

It is also well known that we can use a labelled tableaux method like a decision procedure in modal logic.[1] In order to do so, instead of labelling with truth values, we will use a system of labels of the form 1, 1.1, 1.2, 1.2.1, and so on. The intuitive idea underlying this system is that each label represents a possible world, and the string of numbers and points in which each label consists represents the accessibility relation between worlds. So, the world represented by the label 1.1 is accessible from the world represented by the label 1, 1.2.1 is accessible from 1.2, and so on.

Technically, we define the set of labels in a recursive way:

(a) 1 is a label.
(b) If σ is a label, then $\sigma.n$ is a label (for $n \geqslant 1$)

We will say that a label σ is a simple extension of a label τ if and only if (from now on, iff) σ is of the form $\tau.n$ (for $n \in \mathbb{N}$); and that σ is an extension of τ iff σ is a simple extension of τ, or the simple extension of a simple extension of τ, and so on. Finally, we will say that the world represented by the label σ is reachable from the world represented by the label τ iff σ is an extension of τ.

In order to deal with different systems of basic modal logic, the rules are divided into two kinds that we call *common rules* and *inheritance rules*. The former are the

[1] Q.v., for example, Goré (1999) or Fitting (1983).

same for all systems of the same kind; e.g., for all systems of alethic modal logic. The latter are different for different systems; e.g.: S4 and S5.

The rules for the propositional operators are the usual, with a label which remains invariable (see Table 11.1).

Table 11.1 Common rules for propositional operators

$$\mathbf{R}\wedge:\ \frac{\sigma :: \alpha \wedge \beta}{\begin{array}{c}\sigma :: \alpha\\ \sigma :: \beta\end{array}} \qquad \mathbf{R}\neg\wedge:\ \frac{\sigma :: \neg(\alpha \wedge \beta)}{\sigma :: \neg\alpha \mid \sigma :: \neg\beta}$$

$$\mathbf{R}\vee:\ \frac{\sigma :: \alpha \vee \beta}{\sigma :: \alpha \mid \sigma :: \beta} \qquad \mathbf{R}\neg\vee:\ \frac{\sigma :: \neg(\alpha \vee \beta)}{\begin{array}{c}\sigma :: \neg\alpha\\ \sigma :: \neg\beta\end{array}}$$

$$\mathbf{R}\rightarrow:\ \frac{\sigma :: \alpha \rightarrow \beta}{\sigma :: \neg\alpha \mid \sigma :: \beta} \qquad \mathbf{R}\neg\rightarrow:\ \frac{\sigma :: \neg(\alpha \rightarrow \beta)}{\begin{array}{c}\sigma :: \alpha\\ \sigma :: \neg\beta\end{array}}$$

$$\mathbf{R}\neg:\ \frac{\sigma :: \neg\neg\alpha}{\sigma :: \alpha}$$

The rule for the operator $\Diamond$ is the only one which creates a new label:

$$\mathbf{R}\Diamond:\ \frac{\sigma :: \Diamond\alpha}{\sigma.n :: \alpha}$$

(Where n is the first positive integer such that $\sigma.n$ is new in the branch.)

This rule, together with the rules for the operator $\Box$, may give rise to an infinite branch. We can avoid this situation using the following restriction:

> Except if $\tau :: \alpha$ appears in the branch and σ is reachable from τ. In that case, the rule is considered as applied and the formula is marked.[2]

We arrive now to the rule of the operator $\Box$. This operator, typically, has a different behaviour in different modal systems. If we are dealing, for example, with alethic modal logic, we have to consider the accessibility relation to have some properties besides normality, given by the axiom K. In all these systems, the accessibility relation is also reflexive; the adequate rule is therefore:

$$\mathbf{R}\Box\ \mathbf{refl}.:\ \frac{\sigma :: \Box\alpha}{\sigma :: \alpha}$$

If we are dealing, on the contrary, with deontic modal logic, we have to consider the accessibility relation as serial, but not reflexive. Then the rule is:

$$\mathbf{R}\Box\ \mathbf{ser}.:\ \frac{\sigma :: \Box\alpha}{\sigma :: \Diamond\alpha}$$

[2]That a formula is marked indicates that the corresponding rule has already been applied.

We have talked above about the rules that we call *inheritance rules*. These rules establish the difference between systems and, therefore, depend on the properties of the corresponding accessibility relation. In the most basic case, when the accessibility relation has no more properties than reflexivity or seriality (e.g.: in systems T or KD), the rule is:

$$\textbf{IRT} : \frac{\sigma :: \Box\alpha}{\sigma.n :: \alpha}$$

We can introduce additional properties replacing the preceding rule with a stronger one. For example, if we want the accessibility relation to be *euclidean* (systems S4 or KD4), the rule is:

$$\textbf{IRS4} : \frac{\sigma :: \Box\alpha}{\sigma.n :: \Box\alpha}$$

If we also desire transitivity (systems S5 or KD45), we need this stronger rule[3]:

$$\textbf{IRS5} : \frac{\sigma :: \Box\alpha}{\sigma.n :: \Box\alpha}$$

The most interesting feature of this method is probably what has sometimes been called *modularity*: with the adequate combination of these rules we can create tableau methods for all basic systems of modal logic, such as systems T, S4 and S5 of alethic modal logic and, likewise, systems KD, KD4 and KD45 of deontic modal logic. We can prove all these methods to be sound and complete.[4] Our purpose, in this paper, is to extend these methods to some kinds of multimodal logic, such as multi-agent epistemic (or doxastic) logic and temporal logic.

11.3 Multi-agent Systems

The easier extension of modal logic is the case in which we have a certain number of modal operators of the same kind. This is the case for multi-agent modal logics, the best known of which are epistemic and doxastic logics.[5]

[3]From now on, the double line expresses that the position of the antecedent and the consequent of the rule are interchangeable.

[4]The scheme of this proof is the same in all the cases: to prove that the method is sound we show how to build a model based on an open branch of the tableau and later we prove that this model satisfies the formula. Completeness is proved by induction over the length of the formula.

[5]Q.v.: Fagin et al. (1995).

In this kind of logic we deal with a set $\mathcal{A}$ of agents. Given this set, we introduce an operator $\Box_{a_i}$ and its dual $\Diamond_{a_i}$ for each agent $a_i \in \mathcal{A}$. Often, we write K_{a_i} (and its dual $\widehat{K}_{a_i}$) when we are talking about epistemic logic and B_{a_i}(and its dual $\widehat{B}_{a_i}$) when we are dealing with doxastic logic. When possible, and for the sake of generality, in this paper we shall use $\Box_{a_i}$ and $\Diamond_{a_i}$.

To adapt the tableau method to multi-agent modal logic we only have to adapt the form of the labels:

Given a set $\mathcal{A}$ of agents:

(a) 1 is a label.
(b) If σ is a label, then $\sigma.a_i n$ is a label too (for $n \geq 1$ and $a_i \in \mathcal{A}$).

The intuitive idea underlying this notation is that $\sigma.a_i n$ represents a new possible world accessible from the old one(σ) for the agent a_i.

Regarding to the rules, they are really the same that in the basic case. We only have to adapt them to the new form of the labels. The rule for the operator $\Diamond$ is now:

$$\frac{\sigma :: \Diamond_{a_i}\alpha}{\sigma.a_i n :: \alpha}$$

The rule for the operator $\Box$ depends, like before, on the kind of logic we are dealing with. If we are dealing with epistemic logic, we have to consider the accessibility relation as reflexive, as in the case of alethic modal logic. The rule is therefore:

$$\frac{\sigma :: \Box_{a_i}\alpha}{\sigma :: \alpha}$$

On the other hand, if we deal with a doxastic logic, we have to consider the accessibility relation as serial, but non reflexive, as we have seen in the deontic case. The rule is now:

$$\frac{\sigma :: \Box_{a_i}\alpha}{\sigma :: \Diamond_{a_i}\alpha}$$

With respect to inheritance rules, the new form is:

T_m/KD_m: $\dfrac{\sigma :: \Box_{a_i}\alpha}{\sigma.a_i n :: \alpha}$

$S4_m/KD4_m$: $\dfrac{\sigma :: \Box_{a_i}\alpha}{\sigma.a_i n :: \Box_{a_i}\alpha}$

$S5_m/KD5_m$: $\dfrac{\sigma :: \Box_{a_i}\alpha}{\sigma.a_i n :: \Box_{a_i}\alpha}$

With these easy techniques, we can create tableau methods for all basic systems of epistemic and doxastic modal logic, and we can plausibly extend these procedures

to other kinds of multiagent modal logic. With respect to the systems we have mentioned, we have proved that this method is sound and complete.[6]

11.4 Combining Modalities

In the preceding section, we have dealt with systems where a modal operator for each individual of the group of agents is at our disposal. These systems are usually called "multi-agent systems". But it might be interesting to combine different kinds of modalities, in such a way that we can speak, for example, about the knowledge and the beliefs of the agents in a group, having then a doxastic-epistemic system. In the same way, we can combine deontic and epistemic modalities, and so on.

In order to do so, we have to consider different kinds of accessibility relations and represent them by using the labels. For example, if we want to combine epistemic and doxastic operators, the rules are[7]:

Knowledge $\dfrac{\sigma :: \widehat{K}_{a_i}\alpha}{\sigma.Ea_in :: \alpha}$

(where $\sigma.Ea_in$ is new in the branch)

Belief $\dfrac{\sigma :: \widehat{B}_{a_i}\alpha}{\sigma Da_in :: \alpha}$

(where $\sigma.Da_in$ is new in the branch)

In these rules, the label $\sigma.Ea_in$ represents an epistemic alternative to σ; and σDa_in represents a doxastic alternative to σ.

In regards to the inheritance rules, we have to consider the relations between different kinds of modalities. Depending on the relations we want to accept, we have to change the rules. For example, if we accept that $K_{a_i}\varphi \rightarrow B_{a_i}\varphi$,[8] we have to modify the inheritance rule for K (we have given the example for S4):

$$\frac{\sigma :: K_{a_i}\alpha}{\sigma.Xa_in :: \alpha}$$

(Where X may be E or D)

[6] Q.v.:Gómez-Caminero Parejo (2011)

[7] Since we have to distinguish between epistemic and doxastic operators we can not use $\Box_{a_i}$ and $\Diamond_{a_i}$. We use K_{a_i} and B_{a_i}, and its duals, instead.

[8] A stronger alternative is $K_{a_i}\varphi \rightarrow B_{a_i}K_{a_i}\varphi$ (Hintikka 1962).

11.5 Knowledge in a Group of Agents

We can extend our multi-agent epistemic or doxastic logic (and perhaps a deontic logics, or other modal logics) with operators which try to capture stronger concepts related to the Knowledge (or belief, etc.) of agents which interact in a group.[9] The most usual ones are E and C.[10]

The operator E express the idea that all agents in a group know something. $E\varphi$ (read "everybody knows that φ") is equivalent to $\Box_{a_1}\varphi \wedge \Box_{a_2}\varphi \wedge \cdots$ (for any $a_i \in \mathcal{A}$).

In order to work with tableaux, it is interesting to have at our disposal the dual of the operator E, $\widehat{E}$. $\widehat{E}\varphi$, means *for at least one agent it is possible that* φ. It can be defined as $\Diamond_{a_1}\varphi \vee \Diamond_{a_2}\varphi \vee \cdots$ (for any $a_i \in \mathcal{A}$).

The operator C is stronger than the operator E. Intuitively speaking, $C\varphi$ means "it is common knowledge (belief) that φ". It can be intuitively understood as the infinite conjunction $E\varphi \wedge EE\varphi \wedge EEE\varphi \wedge \cdots$.

Like before, it is interesting to define the dual of the operator C, $\widehat{C}$. $\widehat{C}\varphi$, that we can read "it is compatible with common knowledge (belief) that φ", can be intuitively understood as the infinite disjunction $\widehat{E}\varphi \vee \widehat{E}\widehat{E}\varphi \vee \widehat{E}\widehat{E}\widehat{E}\varphi \vee \cdots$.

How do we deal with this kind of modal operators in order to work with tableaux? The operator E and its dual are not very difficult, given that they can be treated as quantifiers. The rule for E is, therefore:

$$\frac{\sigma :: E\alpha}{\sigma :: \Box_{a_i}\alpha}$$

(For every agent a_i that appears in the branch.)

The rule for $\widehat{E}$ is:

$$\frac{\sigma :: \widehat{E}\alpha}{\sigma :: \Diamond_{a_i}\alpha}$$

(Where agent a_i is new in the branch.)

With respect to the operator C, it is not very difficult either. In fact, it is dealt with in the same way as the operator $\Box$, the only difference lies in the inheritance rules. With respect to the common rules, depending on whether we interpret it as common knowledge or common belief, they are:

Knowledge:

$$\frac{\sigma :: C\alpha}{\sigma :: \alpha}$$

[9] Q.v.: Fagin et al. (1995).

[10] It is also usual to introduce the operator of distributed knowledge D. This notation, it should be noted, is independent from the use of the same capital letters to name epistemic or doxastic alternative relations.

Belief:

$$\frac{\sigma :: C\alpha}{\sigma :: \widehat{C}\alpha}$$

The inheritance rule for the system T is:

$$\frac{\sigma :: C\alpha}{\sigma.a_i n :: C\alpha}$$

whereas the rule for the system S5 is:

$$\frac{\sigma :: C\alpha}{\sigma.a_i n :: C\alpha}$$

The rule for the system S4 is the same as the one for S5, but we have to consider the cases where we have applied the restriction of the rule R$\Diamond$. However, we are not going to explain the details here.

We are arriving now to the difficult point of this section, the operator $\widehat{C}$. But before, let us speak about the DB-tableaux. DB-tableau are modified tableaux for first order logic.[11] With the DB-tableaux we can deal with formulas of the form $\forall x \exists y \varphi(x, y)$. In this modified method, the standard rule

$$\frac{\exists x \varphi}{\varphi\,(k_{n+1}/x)}$$

(where k_n is the last constant that appears in the branch)

is replaced with:

$$\frac{\exists x \varphi}{\varphi\,(k_1/x) \mid \cdots \mid \varphi\,(k_n/x) \mid \varphi\,(k_{n+1}/x)}$$

In this way, if the formula has a finite model, the method finds it in a finite number of steps, although the tableau becomes infinite when the formula does not have a finite model. For example, in the case of the formula $\forall x \exists y R(x, y)$, the method gives us a first open branch

$$\Phi = \{\forall x \exists y R(x, y), \exists y R(a_1, y), R(a_1, a_1)\},$$

then a model with a unique individual in its domain can be defined: domain $\mathfrak{D} = \{\mathbf{a}\}$, and the interpretation function $\mathfrak{I}$ such that

[11] Used in Nepomuceno-Fernández (1999), the method was proposed independently by Díaz Estévez and Boolos.

$$\mathfrak{I}(a_1) = \mathbf{a} \text{ and } \mathfrak{I}(R) = \{\langle \mathbf{a}, \mathbf{a} \rangle\}.$$

We adapt an infinitary version of the previous rule to deal with the operator $\widehat{C}$. We take advantage of the intuitive equivalence between $\widehat{C}$ and the infinite disjunction

$$\widehat{E}\varphi \vee \widehat{E}\widehat{E}\varphi \vee \widehat{E}\widehat{E}\widehat{E}\varphi \ldots$$

Of course, if we find a model for an element of the disjunction, we have found a model for the whole formula. The rule is therefore:

$$\frac{\sigma :: \widehat{C}\alpha}{\sigma :: \widehat{E}\alpha \mid \sigma :: \widehat{E}\widehat{E}\alpha \mid \cdots}$$

(Until we find an open branch.)

As before, if the formula has a model we will find it in a finite number of steps; if not, the tableau becomes infinite in the sense that it has an infinite number of branches (although all of them are eventually closed). We can prove that this method is sound and complete.

11.6 Temporal Logic

We can interpret $\Box$ and $\Diamond$ as temporal operators. In this interpretation, $\Box\varphi$ means that φ is always true (now and in the future) and $\Diamond\varphi$ means that φ is eventually true (now or at some point in the future).

It is also common to introduce two more operators: $\bigcirc$ and U[12] (we are not going to deal with branching-time operators). $\bigcirc\varphi$ means intuitively "at the next moment, φ" whereas $\varphi U \psi$ means "φ until ψ".

Really, it is more common to consider $\bigcirc$ and U as primitive operator and to define $\Box$ and $\Diamond$ in this way:

$$\Diamond\varphi =_{def} \top U \varphi$$
$$\Box\varphi =_{def} \varphi U \bot$$

If we want to modify our tableau method for dealing with temporal logic, we have to introduce the following changes:

Regarding the labels, each one will be a number $t \in \mathbb{N}$ which represents a moment of time (we are dealing with discrete time).

[12]The operator U was first introduced by Kamp (1968). A good introduction to temporal logic is Gabbay et al. (1994).

The rules for $\Box$ are similar to the previous cases. The common rule:

$$\frac{t :: \Box\alpha}{t :: \alpha}$$

And the inheritance rule:

$$\frac{t :: \Box\alpha}{t' :: \Box\alpha}$$

(For any label $t' > t$ that appears in the branch)

The rules for $\bigcirc$ and its negation are:

$$\frac{t :: \bigcirc\alpha}{t+1 :: \alpha}$$

$$\frac{t :: \neg \bigcirc \alpha}{t+1 :: \neg\alpha}$$

Until here everything that at we have done is very similar to the previous case, but now, for dealing with the operators $\Diamond$ and U (and their negations) we have to introduce what we call *recursive rules.*.

Recursive rules are of the form

$$\frac{A}{SC \parallel RC}$$

where SC is the *stop condition* and RC is the *recursive clause*.

The idea is to test each moment of time until we find a model which satisfies our formula. So, if SC gives rise to an open branch, we have found the model and therefore we have finished the application of the rule. On the other hand, if SC gives rise to a closed branch, we have to apply RC, which makes us apply the rule again at the next moment of time. Once again, if the formula has a finite model, we can find it in a finite number of steps; if the formula does not have a model, the tableau becomes infinite.

The easiest rule is the one for $\Diamond$:

$$\frac{t :: \Diamond\alpha}{t :: \alpha \parallel t :: \bigcirc\Diamond\alpha}$$

The rules for U and its negation are more complicated, but essentially based on the same idea:

$$\frac{\alpha U\beta}{t :: \beta \;\Big\|\; \begin{array}{c} t :: \alpha \\ t :: \bigcirc(\alpha U\beta) \end{array}}$$

$$\frac{t :: \neg(\alpha U \beta)}{\begin{array}{c|c} t :: \neg\alpha & t :: \alpha \\ t :: \neg\beta & t :: \neg\beta \\ & t :: \bigcirc\neg(\alpha U \beta) \end{array}}$$

11.7 Conclusions

We have seen that we can present labelled tableau methods for various different systems of alethic modal logic. In this calculus, we use two kinds of rules, the so-called *common rules* and the so-called *inheritance rules*. The former are common to all systems of the same kind, the latter express the properties of the accessibility relation and, therefore, constitute the difference between systems. We can prove that these methods are sound and complete.

We can also extend these methods to multi-modal logics using more complicated labels and rules. We have presented here, although only as a first attempt, the rules for the most usual systems of multi-agent epistemic and doxastict logic.

The situation becomes more difficult when we introduce in our logic infinitary operators, such as the operator C and its dual. With this purpose, and this is a more innovative idea, we propose an infinitary version of what we call a DB-Tableau, originally introduced to deal with formulas of the form $\forall x \exists y \varphi(x, y)$. If the formula has the so-called finite model property,[13] we will find it with this calculus in a finite number of steps; if not, the tableau becomes infinite in the sense that it has an infinite number of branches.

Finally, with the goal of dealing with temporal operators we introduce recursive rules. The easiest case is the rule for the operator $\Diamond$. The intuition underlying this rule is very simple: since we are looking for a model for an expression of the form "eventually φ", we check the model in which φ is true just now. If this possibility ends up being impossible, we have to accept that in the next moment of time is true that φ will eventually be true. In this case, again, when the formula does not have a model, the tableau becomes infinite.

In all of these cases we have proven that the method is sound and complete. We claim that this kind of methods may be extended to other systems of multi-modal logics.

[13] A formula has the finite model property when it is verified that if the formula has a model, then it has a finite model.

References

Fagin, R., Halpern, J.Y., Moses, Y., Vardy, M.Y.: Reasoning About Knowledge. MIT, Cambridge (1995)

Fitting, M.: Proof Methods for Modal and Intuitionistic Logics. Synthese Library, vol. 169. D. Reidel, Dordrecht (1983)

Gabbay, D.M., Hodkinson, I., Reynolds, M.: Temporal Logic: Mathematical Foundations and Computational Aspects, vol. 1. Clarendon, Oxford (1994)

Gómez-Caminero Parejo, E.F.: Tablas Semánticas para Lógica Epistémica. Fénix Editora, Sevilla (2011)

Goré, R.: Tableau methods for modal and temporal logics. In: D'Agostino, M., Gabbay, D.M., Hähnle, R., Possega, J. (eds.) Handbook of Tableau Methods. Kluwer Academic, Dordrecht (1999)

Hintikka, J.: Knowledge and Belief. Cornell University Press, Cornell (1962)

Kamp, J.A.W.: Tense Logic and the Theory of Linear Order. Ph.D. thesis, University of California, Los Angeles (1968)

Nepomuceno-Fernández, A.: Tablas semánticas y metalógica (El caso de la lógica de segundo orden). Crítica **XXXI**(93), 21–47 (1999)

Part III
Argumentation, Conversation and Meaning in Context

Chapter 12
Irony as a Visual Argument

Silvia Martínez Fabregat

Abstract Argumentation fields are extraordinarily varied. Depending on the area in which we move, our argumentative strategies should be appropriate for achieving the greatest success. The strength of a good argumentation must remain meaningfully in an argument developed in a logically valid way and rhetorically embellished, obtaining as a result a persuaded audience who consequently accept it.

Irony, as a rhetorical trope of language, not only embellishes the argument, but it can also be a particularly persuasive argument itself. The ironic argument has some characteristic features such as its dependence on an active audience ready to interpret it, or its proximity to humor, which outlines a characteristic way of approaching the world of the ironic speaker. We will show how irony works within the written speech using Joan Fuster's aphorism as an example; and then, we will explore the possibilities of ironic argumentation in the visual field through one of Banksy's paintings.

Keywords Argumentative strategies • Irony • Rhetoric argument • Visual irony • Joan Fuster • Banksy

12.1 Rethoric Inside Argumentation

Argumentation is the base of our social life. All our relationships need the communicative exchange to be possible. In order to be successful in the dialectic process, we usually employ rhetorical strategies to persuade and eventually we are able to get the support of the audience. The power of the argumentative strategies was well-known by the classics. Aristotle's *Rhetoric*—which brings together most of the former rhetorical theories used by great orators such as Gorgias, the sophist—shows us the importance of the orator's *ethos*, the value of a painstaking *elocutio*, the utility of knowing about our audience's passions and characters as well as providing the hearers with a formally valid argument.

S.M. Fabregat (✉)
Department of Logic and Philosophy of Science, Universitat de València, Valencia, Spain
e-mail: silvia4957@gmail.com

J. Redmond et al. (eds.), *Epistemology, Knowledge and the Impact of Interaction*,
Logic, Epistemology, and the Unity of Science 38, DOI 10.1007/978-3-319-26506-3_12

However, over the centuries, rhetoric was secluded of the argumentation *corpus* because it was considered a tricky technique. Logic outranked it when in the nineteenth century the analytical turn came on philosophic scene. Rhetoric was considered definitely an ornamental issue and the formal expression of the argument was the main way to elucidate its validity, rhetoric could only complicate the task (Toulmin 2003a, p. 88). If Toulmin was right, it would be possible to completely represent our argument by a formal model, despite it not including the rhetorical strategies—which are close to pragmatic and are highly difficult to formalize.

Stephen Toulmin tried to accomplish that task with the so-called Toulminian model of argumentation. That formal outline traces the structure of our arguments attending to its warrants, possible rebuttals, etc. The outline appearance and the syllogism are alike. The English philosopher keeps that in mind as a point of reference, but considers the syllogism too ambiguous to be useful in a precise argument analysis (*Ibid*. p. 100ff). His proposal tries to widen the syllogistic frame explaining the gloomiest aspects.

Nevertheless, although it is very useful to unravel the formal structure and to assess its logic validity, it is totally insufficient to include the rhetoric tropes by which claims can be expressed. And it will not matter if we accept that rhetoric only means aesthetic. If we agree, as we want to show, that rhetorical tropes in general and irony in particular have argumentative value by themselves, we will conclude that rhetoric is not merely a decorative matter. In fact, a formal model such as Toulmin's is not enough to represent the complexity of our arguments. Integrating the pragmatic dimension in the formal sketch is a good way to include rhetoric and the majority of non-literal figures of our natural language. The computerization approach is working in that way trying to add the pragmatic elements to the formal perspective in order to construct an algorithm capable to create or detect, or both, the meaning of non-literal expressions as irony (cf. Reyes et al. 2012; Utsumi 1996). And also the pragma-dialectical approach, defended mainly by F. H. van Eemeren and R. Grootendorst, recently has demonstrated that rhetoric means more than decoration and it is absolutely attached to dialectical exchange (van Eemeren 2010).

12.2 Rhetoric's Argumentative Value

We understand rhetoric as *the spoon full of sugar* which helps the *logos go down*. Sugar is in this case, the set of multiple maneuvers which make the argument as attractive as it is possible to the audience. The uses of rhetorical tropes allude to the different ways that the speakers have to present their arguments depending on the argumentation field where they are working (cf. Toulmin 2003a, p. 11ff), the potential audience that they imagine (cf. Perelman and Olbrechts-Tyteca 2006, p. 55ff) or their argumentative goals (van Eemeren 2010, p. 36ff). The selection of a rhetoric strategy instead of any other, defines the speaker as well as his argumentation.

Analytic utterances rely on universal validity, but the rhetorical strategy which we employ to transmit them is important too in order to persuade de audience.

Speech figures appear in pretty diverse fields from poetry (metaphor, synesthesia, symbol . . . are used to express feelings and sensations which are difficult to describe literally), to science—as J. Fahnestock points out, scientific fields turn to, for instance, metaphor or antimetabole, in order to explain sentences or parts of a theory which could not be an object of demonstration, as for example Newton's third law (1999, p. 140ff.). Even in the most analytic fields, we find rhetorical resources,[1] because all our different choices to express an argument are rhetorical strategies by which we try to affect our audience. Rhetorical figures are useful to clarify gloomy concepts—as we shall prove with "A quasi-political Explanation of the Higgs Boson", David Miller's well-known allegoric explanation of Higgs Bosson[2]—or to name new hi-tech things establishing a similarity to a known one (Black 1962, p. 33). In fact, they are attached to our natural language and we are constantly using them (cf. Lakoff and Johnson 1980), in a way that they stay with us in our new kinds of expression.[3] And, obviously in the substantive utterances rhetoric is also paramount (cf. Toulmin 2003b, p. 37ff) because, when related to *probable matters*, rhetoric is necessary to get the agreement of the audience. That essential position of rhetoric maneuverings is something pointed out by van Eemeren who suggests that the participants involved in a critical discussion want to achieve dialectical objectives in each discussion stage; but simultaneously, they realize analogue rhetorical aims. Hence, in each critical discussion stage there is a rhetorical goal that corresponds with the dialectical goal (van Eemeren 2010 p. 43).

The speakers should know what kind of words are suitable for their audience, what kind of feelings should be raised in each step; what should be explained and how and what it is preferable to hide in order to realize their rhetorical aim which is be persuasive. The ironic strategy is one of the possible means available to achieve that goal. However, in the same way we can use a mobile phone without knowing how it works; we can use the ironic trope in our natural communication without being able to give a complete answer about what it is, how it works, or why we understand its meaning. So let us follow by outlining a general definition of the trope.

[1] When someone tries to explain to another why *sin* of 90° is 0, they have different choices to accomplish it. For instance, I could give to my audience a visual argument using a goniometric circumference or, if I considered that my audience has enough mathematical knowledge, I would show the trigonometrically ratio which demonstrates that if $sin \propto = \frac{opposite\ leg}{hihypotenuse}$, then $sin\, 90° = \frac{1}{1}$.

[2] Available online at http://www.hep.ucl.ac.uk/~djm/higgsa.html [Access May 24, 2013].

[3] New Mass Media and Social Networks are changing our communication system. A few years ago we could easily separate oral from written expression but nowadays, the computer language has been creating a third space which brings together characteristics from the two former ones. For instance, a renowned microbloging service as *Twitter* with more than 500,000,000 users on 2013 (http://www.statisticbrain.com/twitter-statistics/, access May, 24, 2013) and available in the whole world, encourages people to be concise and express a lot in a few characters. These linguistics limits motivate people to use non-literal expressions to convey secondary meanings. This strategy serves to widen the accurate sense of our *tweets*.

12.3 Defining Irony

The speaker, who chooses irony as an argumentative strategy, is discovering his mental description of the communicative act. The ironic trope has been defined in many different ways from diverse disciplines such as literature, linguistic, philosophy or even esthetic. All of them agree in saying that an ironic utterance is the *use of words to convey a meaning that is the opposite of their literal or actual meaning.* But that usual definition seems not wide enough to comprise all the uses of the figure. That abstract definition has been changed to become a more precise proposal.

It is true that the sense of contradiction between two dimensions—an expressed one and a non-expressed one—remains as an essential feature of irony. But it does not explain most of its appearances (Utsumi 1996, p. 2). Other explanations of its constitution understand that trope focused in the way that it happens. The audience must suppose the non-expressed meaning of an utterance starting from the expressed one which usually is different to it—and it is not necessary to be exactly its inversion. The use-mention theory by D. Sperber and D. Wilson attends to this particularity. These authors conceive the communication as an act where it is essential to take into account implicit inferences. These inferences contain the meaning of the speaker's speech and the hearer's need to notice them to understand it completely. The speaker must give his audience some clues to comprehend the real implicit meaning and these are on the words of the message that is transmitted by the speaker, and also in the context where the communication is taking place (Cf. Sperber and Wilson 1981). Irony is understood because the hearers get the meaning that is mentioned but not expressed.

From another point of view, Paul Grice's meaning theory proposes that irony appears when a *conversational maxim* is broken. This author considers firstly the inseparable relation between dialectic and the context where it occurs. The spatio-temporal situation which is shared by the interlocutors confers the meaning to the words that they use. Grice's description of the conversational model shows some unavoidable series of principles which make possible the communicative process. The main one is the *cooperation principle* (cf. Grice 1989, p. 26ff). The rest of conversational maxims consist of specifications of this principle. Overflowing these maxims does not mean going into fallacy land. It could be highlighting the presence of a rhetorical trope such as irony, as we could read from R.N. Norrick.[4]

[4]"Irony as a violation of his so-called 'conversational maxims'. The maxims represent rules for logical, expeditious talk which speakers act as if they were following. They consist in rules like be brief, be orderly, be relevant, and so on. Apparent violations lead listeners to search for an interpretation in line with the overarching 'Cooperative Principle' as follows: if you say Nice tie but I know you do not like paisley ties, I will construct an interpretation for your utterance assuming you intended something special in violating the maxim of quality, namely, that you want me to recognize that you are following a convention whereby were speaking ironically and do not in fact like the tie at all, especially since irony always reflects a hostile or derogatory judgment." (Norrick 1993, p. 155)

However, these definitions are insufficient, as A. Utsumi has pointed out. The reason is that no one independently could give an answer to the three essential questions to define that trope: "what properties distinguish irony from non-ironic utterances? How do hearers recognize utterances to be ironic? And what do ironic utterances convey to hearers?" (1996, p. 1). Breaking a conversational maxim is not enough to explain how each kind of irony works. In fact, some of them can be communicated among expressions which do not break any maxim, and it seems that the concept of 'mention' is too vague to provide an explanation to these three questions. From the computational theory, Utsumi proposes a new model which combines and spreads some of the previous theories. This theory has been working in a definition in order to compose a functional algorithm to recognize ironic utterances. Although that theory is focused on verbal and situational irony, great results are obtained in that area. Utsumi is one of the main authors who are exploring that dimension. His theory considers that an ironic utterance implicitly displays *ironic environment*, which is characterized by three special properties for being considered as ironic. Firstly, *allusion*: the concept of allusion proposed by Utsumi lies in Kumon-Nakamura's conception, according to which "ironic utterances allude to a failed expectation and violate one of the felicity conditions for well-formed speech acts" (Utsumi 1996, p. 31), Secondly, *pragmatic insincerity*. Irony intentionally violates pragmatic principles (*Ibid*. p. 32). And, thirdly, *emotional attitude*, the speakers communicate their emotional attitude through lots of different signals (*Ibid*. p. 33). These properties are important to express how the expectation expressed by the utterance fails.

Using that definition as a reference, we will explain the argumentative power of two different arguments which use irony as a weapon in the dialectical field.

12.4 Irony in Words

An essay is a non-specialized or exhaustive prose it is "an incitement to conversation"[5] that's the reason because essayists picks up literary figures close to humor and typical of the dialogue. The use of these tropes brings the author close to the reader. It alludes to a common horizon of meanings that they are watching while apparently their attention focuses on another point. That active involvement of the audience and the understanding among them is possible only by a trope such as irony. For that reason, we have selected an aphorism from *Diccionari per a ociosos*, written by the well-known essayist Joan Fuster to analyse how irony work as an argumentative strategy.

> I do not understand who said that they underestimated money. It takes so much hard work to earn it![6]

[5] "L'assaig és (. . .) una incitació a la conversa." (Fuster 1991, p. 9).

[6] "No entenc aquells qui diuen que menyspreen els diners. Costen tant de guanyar!" (Fuster 2009, p. 39).

If we attempt a natural understanding of the utterance, we will need to admit that the author is giving us a lesson about the value of things and a piece of sense of humor taking money as an unimportant issue. Being an aphorism it has a powerful feature which it also shares with irony: it transmits a lot with only a few words. While the claim is only the tip of the iceberg, the second meaning, which is alluded, looks like the huge mass of ice under the surface.

The clues that build the ironic context and conduct us to the second meaning appear as follows:

1. The expectation of the speaker fails because exist a(n intentional) misunderstanding between the citation of the first sentence and the answer given in the second one. The interpretation of 'underestimated' alludes to different meanings of 'money worth.'
2. The maxim of quality is broken because the speaker is not trustworthy. Fuster knows that people who underestimate money are not referring to the cost of earning it but a deeper meaning of the valuable things in life. His forced ridiculous attitude is looking to encourage a reaction from the audience.
3. The emotional attitude of the speaker gives us the last clue: it is strange that someone like Fuster doesn't understand the position of who says underestimate money. This proclamation of a *known error,* in W. Booth's terminology (1975, p. 57), will be only allowable if it points out an ironic meaning.

Why did he choose irony as a weapon? Maybe a social request can give us an interesting reason. We must remember that the vast majority of his work was produced and published during the Spanish dictatorship, when censorship kept an eye on whole words written in the country. In that context it was essential to be careful about what to say and what it was necessary to keep hidden. Irony offers a chance to say without being exposed. In other words, as Kierkegaard wrote, who use irony is *free negatively* because at one and the same time what it is thought is not the same as what is said; the speaker is free from the hearers and from himself (2000, pp. 287–288).

But Joan Fuster[7] may have a different answer for us. He did not want that his sharp style will be considered as a wanton resource. He understood his task as a criticism which had to serve as a corrector to absurd situations. And that is the point. From this view, the world is a place where ridiculous events are happening whose real nature someone has to point out and uncover. As in *The Emperor's new suit*, his irony discovers the nakedness of those who do not want to be called into question.

Aristotle said that "Irony better befits a gentleman than buffoonery" (*Rhetoric* 1419B5), that's why Fuster, as elegantly as an ironic comment can be, shows the hidden aspects of an untruthful reality. In that case, the trope is used to pass over a politically correct utterance, and it is oriented directly to an audience willing to

[7]From http://vimeo.com/46346382 (Access May, 21, 2013).

interpret the second meanings of the words. Its approach to humor is also useful to report the absurd nature of some established situations.

Finally, we do not have to forget that essays and conversations are alike, and irony gives to the audience the possibility of interacting, interpreting and supporting the argument. These features make irony an invitational resource that is tremendously persuasive.

12.5 The Message in the Wall

The England-based graffiti artist known as Banksy is bordering on the limits of legality when using Bristol's streets to express his ideas. And he is shaking up the art scene with his distinctive stenciling technique and his unorthodox way of developing his artistic career. But most attractive to all of us is the ironic tone of his work.

Maybe its consideration as a piece of art is still arguable, but there is no doubt about the media impact of his work. When a new graffiti signed by the characteristic aerosol lines of Banksy appears on a wall or on a bridge, he switches on the process of communication and pedestrians are his audience. Sometimes his pictures are joined to short written messages, and although that part could be a verbal argument, we intend to focus in the strictly visual part.

In the last few years the existence of visual argumentation has become socially accepted. Nowadays, images move the world and bring us powerful messages which, in a more evidential way than words, need to be interpreted. Images can have an argumentative role in three different ways, as L. Groarke points out (2002). They can serve as a backdrop of an argumentation without real relevance. But that use could be significant if images were intentionally persuasive, in a way that they attract the audience's attention to the verbal argument. In that case they will be considered as a 'visual flag'. However, the stronger argumentative value of an image is represented by the third type of images which can be interpreted as a speech act. We will not discuss whether images could be arguments.[8] We want to show how irony can be expressed by means like an image, particularly in the street art of Banksy.

Napalm (Can't Beat The Feeling) (Fig. 12.1[9]) shows Ronald MacDonald and Mickey Mouse, two characters associated with positive feelings—Fun, happiness... Holding their hands, between them, is a Vietnamese naked little

[8]This was successfully defended in a pragma-dialectical setting. Groarke (2002) alludes to five pragma-dialectical principles, that are accomplished by argumentative images (van Eemeren and Grootendorst 1992, pp. 49–55): They are understandable; through metaphorical language the composition elements can be read and interpreted because have an internal sense; they are related to the social and historical moment, in fact, an external point of view can explain its meaning; and an image could be enough to solve a conflict.

[9]From http://cincuentamas.tumblr.com/post/14126090078/inspirados-picnic-kibun (Access May 24, 2013).

Fig. 12.1 Napalm (Can't beet the feeling). London, 2004

girl, wracked with pain, running away from napalm which has burned her clothes and now it is burning her skin. We consider this work as an ironic argument because it accomplishes not only the ironic features but it is also argumentative.

So, we can conclude that this Banksy's work is ironic considering the features of the standard definition that we can summarise thus:

1. The image makes allusion to a second non-literal sense. That sense appears when we realise that the speaker's expectation fails because of some elements of the composition.

When the speaker/painter brings together the three characters, he generates according to W. Booth, a *conflict of beliefs* (1975 p. 73), which can only be explained by noticing the irony in the composition. Actually, when we read the elements of the composition—which can be read as an argument—, we can only elucidate its relations by supposing a second meaning which the image makes allusion to. Following Utsumi's definition of allusion, the expectation of the speaker—offering a positive image of the USA values—is broken by the element of the little child. The expectation fails because the assumptions related to the image's symbols are not correlative to a unique and blissful meaning.

2. The author's emotional attitude shows an apparently happy mood when the heart-rending image of Kim Phuc talks about meanness and sadness. That emotional contradiction can be understood from Booth's proposals as the "proclamation of a known error" (1975, p. 57), acting as an indication of the ironic presence.

The author's emotional attitude can be expressed also by the rawness of the subdued colour, empty of cheerful signals and pretty numb to the positive feelings of the outer characters and to the negative of the middle one. On the other hand, joining together Mickey Mouse (which is a symbol of the innocence of childhood, fantasy and joy), Ronald McDonald (Symbol not only of an American way of food and life, but also of an attractive model of modernism, market globalization and capitalism), and Kim Phuc (icon of war horrors in which the USA was involved) shows the difficulties of being glad about the North American conquests and the goodness of its cool way of living, without remembering its incommensurable destruction power.

3. Banksy broke the maxim of quality (Cf. Grice 1989) when he is not truthful in connecting these three elements (Mickey Mouse, Ronald MacDonald and the Napalm's victim) as equals in a context that seems to represent the happiness, in a first, literal view.

Although the picture meaning is explicit, it could be clearer if we pay attention to the title: *Napalm (can't beat the feeling)*. "Can't beat the feeling" is a renowned slogan from Coca-Cola—another USA's brand—which was popularized by an advertisement campaign in 1989. In it, the product was showed as a cheerful drink and the message transmitted was: you can't beat the feeling of happiness and energy when you get that product. But when the word 'napalm' is located just before the sentence, it changes its meaning completely. Banksy twist the meaning and the mood of the words and give the viewers an alternative message which we could understand as: you can't beat the feeling of suffocation and pain caused by the gas. The opposition between de emotions which are transmitted by the sentence is obvious and represent another sign of the presence of irony in this work.

Consequently, attending to the title and to the drawing, and since a painting could constitute an argument and it can be as figurative as verbal language, or even more, we can conclude that interpreting this graffiti as an ironic argument is legitimate. We can classify that irony as a *stable-overt* one because it "require[s] no special act of reconstitution or translation, because [it] *assert* an irony in things or events that the speaker has observed and wants to share." (Booth 1975 p. 236) In that case, Banksy points out the contradiction between the appearance of USA and its actions.

But the interpretation of images is wider than the interpretation of words. They use a code, and some drawings have been associated to traditional meanings as words do (Carrere and Saborit 2000, p. 70ff); but pictures develop feelings and experiences in spectators that may be not contained in the painter's intention (ibid. p. 65). The possibilities of interpretation increase according as we move from a representative art to an abstract one. There is not a unique explanation for an image, despite the author could points out in a particularly direction, maybe through the title. It could be a negative feature related to the use of images as argument, because the author could easily fail in his communicative aim in case the spectator does not understand him. But it is a risk that shares with the oral or written expression. However, on the other hand, it could be a hugely rich mean of communicative exchange because it is not subject to a language which needs to be translated. Pictures use a universal language. Although the references could be more accessible

for some audience than another—for instance, *Napalm (Can't beat the feeling)* could be understandable in a wider sense for a western audience than for people who don't know that iconic symbols—a picture offers a visual support comprehensive for almost every one. In spite of the possible audience don't know who is Mickey Mouse or Kimn Phuc, they could understand the mood showed by the characters and their unbalance. Banksys's graffiti is a very disturbing picture because of the numb sensation oriented to the tearful girl who is led by the hands, in a parade sponsored by Disney and McDonalds. But, despite the audience could identify only an anonymous little girl and two cheerful characters, the dialectical strength of the picture is enormous.

Then, we can conclude that and image could be argumentative and ironical. If a written irony is, as Booth said, "richer than any translation we might attempt into non-ironic language" (1975, p. 6), when we select an image to compose an ironic strategy we would enlarge its meaning wealth.

12.6 Conclusion

In terms of oral or written expression, either in the visual field, rhetoric is substantive to get the aim of persuade the audience. Literally figures are useful resources to achieve that aim. And most of them, as irony does, are malleable enough to be argumentative in many different ways.

We showed how irony works in a Joan Fuster's essay and in a Banksy's artistic work. In all such cases ironic meaning can be described as something that "happens" (Hutcheon 2005, p. 58) from the combination of words or elements taken from the picture. Its presence involves a wider perspective of the point. It is impossible that the literal form keeps the same level of powerful, complicity with the audience—at least, we are sharing a secret—and elegancy, that we discover in irony.

Bibliographical References

Aristotle: Rhetoric. Edited by Lee Honeycutt (Online). Available in http://rhetoric.eserver.org/aristotle/index.html. Access 25 May 2013

Black, M.: Models and Metaphors: Studies in Language and Philosophy. Cornell University Press, Ithaca (1962)

Booth, W.C.: A Rethoric of Irony. The University of Chicago Press, Chicago/London (1975)

Carrere, A., Saborit, J.: Retórica de la pintura. Cátedra, Madrid (2000)

Fahnestock, J.: Rhetorical Figures in Science. Oxford University Press, London (1999)

Fuster, J.: Ser Joan Fuster. Antologia de textos fusterians. Bromera, Alzira (1991)

Fuster, J.: Diccionari per a ociosos. Educaula, Barcelona (2009)

Grice, H.P.: Studies in the Way of Words. Harvard University Press, London (1989)

Groarke, L.: Toward a pragma-dialectics of visual argument. In: van Eemeren, F.H. (ed.) Advances in Pragma-Dialectics, pp. 137–151. Sic Sat/Vale Press, Amsterdam/Newport News (2002)

Hutcheon, L.: Irony's Edge. The Theory and Politics of Irony. Routledge, London (2005)

Kierkegaard, S.: Sobre el concepto de ironía. In: Larrañeta, R., Gonzàlez, D., Saez, B. (eds.) Escritos de Soren Kierkegaard, vol. I. Trotta, Madrid (2000)
Lakoff, G., Johnson, M.: Metaphors We Live By. University of Chicago Press, London (1980)
Norrick, R.N.: Conversational Joking: Humor in Everyday Talk. Indiana University Press, Bloomington (1993)
Perelman, C., Olbrechts-Tyteca, L.: Tratado de la argumentación: la nueva retórica. Gredos, Madrid (2006)
Reyes, A., Rosso, P., Buscaldi, D.: From humor recognition to irony detection: the figurative language of social media. Data Knowl. Eng. **74**, 1–12 (2012)
Sperber, D., Wilson, D.: Irony and the use-mention distinction. In: Cole, P. (ed.) Radical Pragmatics, pp. 295–318. Academic, London (1981)
Toulmin, S.: The Uses of Argument. Cambridge University Press, Cambridge (2003a)
Toulmin, S.: Regreso a la razón. Península, Barcelona (2003b)
Utsumi, A.: Implicit display theory of verbal irony: towards a computational model of irony, In: Proceedings of the International Workshop on Computational Humor (IWCH96), pp. 29–38 (1996)
van Eemeren, F.H., Grootendorst, R.: Argumentation, Communication, and Fallacies: A Pragma-Dialectical Perspective. Lawrence Erlbaum Associates, Hillsdale (1992)
van Eemeren, F.H.: Strategic Maneuvering in Argumentative Discourse: Extending the Pragma-Dialectical Theory of Argumentation. John Benjamins Publishing Co., Amsterdam (2010)

Electronic References

Banksy: Napalm (Can't Beat the Feeling). Available online in http://cincuentamas.tumblr.com/post/14126090078/inspirados-picnic-kibun. Access 24 May 2013 (2004)
Miller, D.J.: A Quasi-Political Explanation of the Higgs Boson. Available on line in http://www.hep.ucl.ac.uk/~djm/higgsa.htm. Access 24 May 2013 (1993)
Roig, M.: Interview to Joan Fuster. In: Personatges (TVE Catalunya). Available on line in http://vimeo.com/46346382. Access 21 May 2013 (1977)
Statistic brain: http://www.statisticbrain.com/twitter-statistics/. Access 24 May 2013

Chapter 13
Ascribing Knowledge to Experts: A Virtue-Contextualist Approach

Sruthi Rothenfluch

Abstract I argue that epistemic contextualism, as conceived by Lewis and DeRose, cannot accommodate knowledge-ascribing behavior in contexts where expert counsel is sought. Narrowly focusing on the subject's epistemic position with respect to p in $\sim p$ possibilities yields the wrong verdict in such cases. To account for our judgments, I propose that contextualists should look to virtue responsibilism, which founds epistemic evaluation both on the mastery of relevant underlying principles and their explicit and implicit application. Such assessment is not measured by S's ability to rule out relevant alternatives or track the truth of p, and for this reason, is not captured by either version of contextualism.

Keywords Expert • Epistemic virtue • Responsibilism • Contextualism

Contextualists maintain that the truth conditions of knowledge ascriptions vary according to certain contextual features, claiming that what shifts from one context to another is the epistemic strength required to know the proposition, cashed out in terms of relevant alternatives and possible worlds. Lewis writes that a subject knows that p in a certain context c if and only if she is able to rule out *relevant* not-p alternatives. DeRose maintains that we are inclined to accept 'S knows p' only if we believe that the subject can track the truth of p in close $\sim p$ worlds. Both, then, assess the truth of knowledge ascriptions by asking whether the subject's epistemic position would be strong enough to allow her to appropriately respond to cases in which $\sim p$. The problem is that this picture does not accommodate our judgments in contexts where expert counsel is sought. When individuals seek expert advice, a non-expert S may very well meet the conditions laid out by DeRose and Lewis and yet not be ascribed knowledge by her interlocutors. This is because in such contexts, we value a firm understanding of unifying principles within the field and insightful deliberation based on such understanding. These qualities cannot be assessed by the more traditional contextualist models offered by Lewis or DeRose. Rather, expert contexts require consideration of the subject's exercise of intellectual

S. Rothenfluch, Ph.D. (✉)
Visiting Instructor, University of Portland, Portland, OR, USA
e-mail: sruthirachel@gmail.com

J. Redmond et al. (eds.), *Epistemology, Knowledge and the Impact of Interaction*,
Logic, Epistemology, and the Unity of Science 38, DOI 10.1007/978-3-319-26506-3_13

virtue. That is, whether knowledge can truly be ascribed to *S* will be determined by whether or not *S* has virtuously arrived at her belief. Admittedly, this will generate a less unified theory because considerations of virtue will emerge only in certain contexts. Why should contextualists accommodate our judgments in these contexts and thereby accept a more fragmented account? One of the main sources of support for contextualism comes from its ability to account for our ordinary knowledge-ascribing behavior without imputing massive and systematic error to competent speakers. Such support, then, is significantly weakened if it excludes our judgments whenever we seek expert advice. Further, my proposal is consistent with the overarching contextualist thesis that interests of conversational participants will determine the sorts of considerations that are relevant to knowledge.

I will begin here by briefly reviewing the tenets of epistemic contextualism, using David Lewis's and Keith DeRose's accounts as my primary sources. While there are other strains of contextualism, I take issue with the general form endorsed by these two, and given their extensive and influential work on the topic, will refer to their approach as traditional contextualism. Next, I discuss a case that challenges this conception and consider ways in which a traditional contextualist might defend their position. I contend, however, that such efforts do not succeed and will draw on virtue responsibilism to propose an emendation to this view.

13.1 Traditional Contextualism

Contextualists maintain that the content or truth conditions of knowledge assertions of the kind 'S knows that p' or 'S doesn't know that p' vary according to the context of the speaker. Certain features of the speaker's context—the interests of conversational participants, stakes involved, etc—determine how good an epistemic position S must be in with regard to p in order to know that p. Thus, according to the theory it is perfectly consistent for a speaker to affirm that S has knowledge that p in one context, and deny that the same subject possessing the same amount of evidence knows p in another context that demands greater epistemic strength with respect to p. The best support for such a theory comes from our linguistic behavior in everyday, non-philosophical contexts. As the ever-growing number of examples[1] in the literature show, the standards of knowledge systematically shift up as the consequences of being wrong become more severe, and the doubts and counter-possibilities considered increase. Keith DeRose's bank case is illustrative: In Bank case A, a man and his wife arrive at the bank to deposit their checks only to find a long queue. The man suggests that they return to deposit their checks tomorrow. His

[1]Take for example, Matt McGrath's train cases in "Evidence, Pragmatics and Justification," *Philosophical Review* 111, no. 1 (2002): 67–94. Stewart Cohen's airport cases in "Contextualism, Skepticism and The Structure of Reasons," *Philosophical Perspectives 13: Epistemology,* ed. James E. Tomberlin (Atascadero: Ridgeway, 1999): 57–89.

wife points out that many banks are closed on Saturdays. But, the man responds, "I was here 2 weeks ago on Saturday, so I know the bank will be open tomorrow". In Bank case B, the couple finds themselves in the same situation, and the man once again suggests coming back the next day. But this time, they have just written a very large check, and face severe penalties if their check is not deposited. His wife raises these points and says, "Banks do change their hours, do you know that the bank will be open tomorrow?" At this point, the husband says, "Well, no, I'd better check to make sure".[2] Note that while the husband's evidence for *p* remained constant, we are inclined to accept the husband's knowledge ascription in the first case as well as his denial in the second.

Before proceeding, there are two important points to note about this example. First, the husband's response in both cases seems right, and if taken at face value, supports a contextualist analysis of knowledge ascriptions.[3] Second, this example is representative of the majority of cases presented in contextualist literature, where epistemic evaluation concerns the subject's epistemic status with respect to a single belief which is itself not founded in, or connected to, a broad understanding of interrelated facts that explain the truth of the belief. In the next section, I will present a different type of example in which knowledge judgments are not based solely on the subject's relation to the particular belief at hand, but involve considerations that have to do with a deep understanding of the subject matter and an excellence in applying this understanding to yield the belief in question. As will be shown below, Lewis's and DeRose's versions of contextualism are not equipped to handle such situations, despite their success in more common cases. To see why, we need to have a better understanding of what it is that shifts from one context to the next, and more importantly, when a subject is said to have met these conditions.

Epistemic contextualists have interpreted the variability of knowledge ascriptions in different ways. According to David Lewis, S knows that *p* if and only if S's evidence eliminates every possibility in which not-*p*. The domain of 'every possibility' is restricted to relevant counter-possibilities determined by specific rules applied to the contexts of both subject and ascriber.[4] For example, according to Lewis's Rule of Actuality, the possibility that actually obtains in the *subject's* context is never properly ignored, while the Rule of Attention dictates that possibilities that are not being ignored in the conversational (ascriber's) context are not properly ignored. These and other rules demarcate the set of possibilities that the subject's evidence must eliminate in order to know that *p* in a given context. What shifts for Lewis,

[2]Keith DeRose, "Contextualism and Knowledge Attributions" *Philosophy and Phenomenological Research* 52, no. 4 (December 1992): 913–929.

[3]Invariantists offer an alternative explanation according to which the standards of knowledge remain fixed, but what is communicated varies from one context to the next. See, for example, Patrick Rysiew, "The Context-Sensitivity of Knowledge Attributions". I will not be discussing this view here as my objective is to present and suggest a possible solution for problems internal to contextualism.

[4]David Lewis, "Elusive Knowledge," in *Skepticism: A Contemporary Reader,* eds. Keith DeRose and Ted Warfield (Oxford: Oxford University Press, 1999): 225.

then, is the set of possibilities or alternatives that may not be properly ignored. The mechanisms responsible for causing such shifts are his rules, which function either to preclude or include certain alternatives as relevant.

How do subjects satisfy these standards? That is, what does it mean to *rule out* alternatives? Dretske maintains that the agent must *know* that the relevant alternative does not obtain:

> In saying that he must be in a position to exclude these possibilities I mean that his evidence or justification for thinking these alternatives are *not* the case must be good enough to say he *knows* that they are not the case.[5]

Dretske's description does not illuminate the notion of ruling out when 'ruling out' is itself used as an analysis of knowledge ascriptions. A more useful description would be one that explains this relation by invoking other features of the subject's epistemic position. For Lewis, a counter-possibility is uneliminated if it does not conflict with the subject's perceptual or memorial evidence. A world *W* is eliminated when a subject's actual experience conflicts with *W*. By 'conflict', Lewis means that the subject does not have the experience. I cannot rule out worlds in which I'm a brain in a vat electro-chemically stimulated to have the experiences of ordinarily life because I would have exactly the same experience and memories in such a world. In contexts where this world is relevant (say where we are reflecting on skeptical possibilities) I will not know this and many ordinary propositions precisely because my experience and memory will be the same in such worlds. On the other hand, in ordinary contexts (say a low stakes situation where no skeptical possibilities have been considered), I know a lot more. For example, in a casual conversation with a colleague, I can truly claim to know that I left my coat in my office because I would not have the memory of hanging my coat on my office door in the range of relevant worlds in which I did not leave my coat in my office. What is important to note here is that whether or not it is true that one knows p in context c is determined by one's epistemic status with respect to p in a circumscribed set of $\sim p$ worlds.

According to DeRose, we are inclined to accept 'S knows that p' when we think that S's belief that p is *sensitive*.

> When it is asserted that some subject S knows (or does not know) some proposition P, the standards for knowledge (the standards or how good an epistemic position one must be in to count as knowing) tend to be raised, if need be to such a level as to require S's belief in that P to be sensitive for it to count as knowledge[6]

DeRose incorporates Nozick's condition of sensitivity by maintaining that when "S knows/does not know that p" is uttered, we will attribute knowledge to S only if we believe that S would abandon her belief that p in the closest not-p worlds. This rule also seems to aptly explain our epistemic position with respect to skeptical

[5]Fred Dretske, "The Pragmatic Dimension of Knowledge," *Philosophical Studies* 40, no. 3. (Nov., 1981): 370–1.

[6]Keith DeRose, "Solving the Skeptical Problem," in Skepticism: A Contemporary Reader, eds. Keith DeRose and Ted Warfield (Oxford: Oxford University Press, 1999): 206.

hypotheses. The assertion that I don't know that I'm not a brain in a vat ratchets the standard of knowledge up to include sensitivity, such that in order to know that I'm not a brain in a vat, I would have to give up that belief in worlds where I am a brain in a vat. Given that such a world is very distant, I can be in a very strong epistemic position and yet not know, since my belief that p will not be sensitive. On the other hand, when we are talking about knowing where I left my coat, my belief appears sensitive. In the closest $\sim p$ world (say, where I left my coat in my car) I would not form the belief that I left my coat in my office. For this reason, I can claim to know that I left my coat in my office.

Both accounts effectively explain variability in our knowledge-ascribing behavior, including our judgments in skeptical contexts. However, they fail to accommodate our knowledge ascriptions and denials in contexts where expert counsel is sought. This is because such judgments hinge not on considerations of sensitivity or ruling out, but on the subject's grasp of principles within the field and her ability to appropriately utilize such understanding. Crucially, the latter is not measured by her ability to respond appropriately to $\sim p$ worlds.

13.2 Expert Scenarios: Diagnosis

PTSD

Upon returning from his deployment to Afghanistan, Joe experiences erratic mood swings, weight loss, nightmares and a lack of interest in normal activities. He makes an appointment with his doctor, but when he arrives finds that Dr. Franklin has suddenly taken leave and a technician has offered to examine him. After a brief interview, the technician asserts that her patient is suffering from PTSD. The solider is, understandably, skeptical about the technician's competence. But the tech assures him that she has watched the doctor on a number of occasions, and has become rather proficient at diagnosis, emphasizing that she knows that he has PTSD. Nevertheless, Joe returns the next day when he is able to see Dr. Franklin, who conducts a very similar interview and confirms that the patient is in fact suffering from PTSD.

The tech's meticulous and consistent observations of the doctor's methods allow her to identify PTSD with success by looking for symptoms that the doctor has found telling in the past. One can imagine, perhaps, a check-off list that the tech has stored in her mind, which generally leads to the correct diagnosis. On the other hand, Joe's decision to consult the doctor appears prudent.

How would a traditional contextualist such as DeRose assess the patient's position? For DeRose, Joe's acceptance of the tech's claim to know depends on considerations of the sensitivity of the tech's belief. We can sharpen the case: suppose that the tech explains her check-off system to Joe, and says something like the following, "So, you see, you exhibit all the classic symptoms of PTSD". The list enables the tech to reliably discriminate PTSD from non-PTSD patients, and in a way that Joe understands. There is little reason for Joe to judge that the technician, using this method, would believe that Joe had PTSD if he did not. In other words, Joe would judge the tech's belief to be sensitive. According to DeRose's brand of

contextualism, then, Joe would ascribe knowledge to the tech. Given that the tech's experiences would be different in relevant $\sim p$ worlds, say because she would not have observed symptoms of PTSD on her list, Lewis would be committed to the same conclusion.

It seems intuitively right, however, for Joe to have made an appointment with the physician the next day, despite the tech's resassurances. Can the contextualist explain this tension? One explanation, consistent with the verdict presented in the preceding paragraph, might be that even though Joe accepts the technician's claim to know his diagnosis, Joe seeks *additional* information, such as the best strategies to cope with the illness, the toll this might take on his friends and family, its impact on his career, etc. Joe's interest in these surrounding issues seems reasonable and likely. But I don't think this explains the situation. Rather, it seems very clear that someone who reaches out to a physician for an explanation of his debilitating symptoms would not and should not ascribe knowledge to a technician. The reader, however, might not share my intuitions in this case, so I will offer two reasons in its support.

First, ascribing knowledge to the technician commits us to the rather implausible view that anyone (with or without medical training) is qualified to diagnose serious illnesses, so long as he or she has *observed* a physician. While this might be conceivable for certain professions that allow new recruits to learn as they go, this practice seems implausible in medicine. We believe that the sort of complexity involved in detecting and treating mental and physiological illness cannot be grasped and mastered through informal observation. The doctor will presumably tap into different reservoirs of information pertaining to his studies in human psychology, current research and his past clinical experience, integrating this information in explicit and tacit ways, to generate an accurate diagnosis. Further, it is not clear that the tech's epistemic situation would improve if he were to ask the doctor to explain his decisions. Studies have shown that experts, particularly where they are making routine decisions use non-reflectively accessible methods.[7] Even if doctors could articulate these processes, merely knowing what they are would not ensure the tech's ability to execute them. This point generalizes to other situations in which expert counsel is sought. Individuals approach specialists precisely because they want someone with a sufficiently deep understanding of the subject matter, someone who understands why and how p. To assess the fairness of a recent Supreme Court decision, for example, journalists will reach out to historians and professors of law, given their familiarity with previous court cases and the law, and the circumstances of the present ruling, all of which equip these experts with a comprehensive understanding of the situation.

Second, taking the technician's self-assessment as accurate discounts the importance of the expert's deliberative process. We appeal to stock brokers, advertising consultants, and political advisors, both because of their expansive grasp of the field

[7]See discussion in Duncan Pritchard, "Virtue Epistemology and the Acquisition of Knowledge," *Philosophical Explorations* 8, no. 3 (September 2005): 229–243.

and their ability to use such understanding in careful, insightful and creative ways to yield answers to our questions. We expect, in other words, that expert deliberation will be distinctive. While these considerations will not be relevant to low-standard cases, say where we are wondering about the hours of a local hardware store, or whether our flight has a lay-over in Chicago, they seem critical to a certain class of high-grade knowledge, where the relevant proposition is based on a complex network of historical or scientific facts, which are then carefully applied to yield solutions and answers for the problem at hand.

A traditional contextualist might contend that I have not correctly applied their theory. In this setting, the range of relevant not-p worlds is wider than I've assumed. Suppose that while the tech is correct that the patient is suffering from PTSD, the patient's symptoms are consistent with a number of different diagnoses. In a close possible world in which the patient did not have PTSD but presented with many of the same symptoms, the doctor, but not the tech, would not hold the belief that the patient suffered from PTSD. An experienced doctor, but not the technician, would be able to draw on his years of clinical practice and medical training to identify certain nuanced symptoms and settle on the right diagnosis. Indeed, the case so described will be one that the traditional contextualist may be able to handle, precisely because the difference between the tech and doctor appears to hinge on the sorts of epistemic differences identified by DeRose and Lewis. However, the case at issue is different because it features two individuals who bear the same relation to the proposition, insofar as both would be able to eliminate all relevant alternatives on account of their evidence and track its truth in relevant $\sim p$ worlds. This is because Joe presents with symptoms that point to a single diagnosis, which allow both the doctor and tech to meet the standards for epistemic strength as defined by traditional contextualism. The problem is that despite meeting such conditions, we have good reason to think that Joe would and should reject the tech's claim to know. What this suggests is that DeRose and Lewis's accounts are inadequate in these contexts because we do not distinguish knowers by considering their epistemic position in $\sim p$ worlds.

13.3 Evaluating Inquiry and Deliberation

What seems to drive our epistemic evaluation in this context (and more generally, any in which individuals seek expert advice) is our interest in the putative knower's broad understanding of the field and her competence in working with the data, which includes her aptitude for threading together relevant information to construct coherent explanations and her innovativeness and success in generating likely hypotheses. To accommodate these judgments, contextualists must shift away from relevant-alternatives and sensitivity frameworks and towards evaluation of the agent's deliberative process. One way to do this is by adverting to the norms of inquiry and deliberation found in discussions of virtue responsibilism, a brand of virtue epistemology. Possessing virtue, according to responsibilists, involves both the learning and mastery of principles within a field and their adept application.

In what follows, I will describe in some detail what this consists in and end by considering some concerns against this approach.[8]

According to Zagzebski, virtue involves two components: motivation that drives an agent to learn the rules and procedures of belief formation accepted by her epistemic community and success in carrying these out.[9] Zagzebski identifies two different stages at which motivation operates: At a foundational level, the subject is driven by an underlying motivation for knowledge, or more broadly, for cognitive contact with reality (where this might also include understanding). This motivation leads the agent to adopt individual motivations distinctive of intellectual virtues such as the motivation to be open-minded in collecting and appraising evidence, fair in evaluating the arguments of others, diligent in recognizing and addressing recalcitrant data, etc. These virtues in turn drive the agent to adopt specialized cognitive skills that are conducive to achieving these aims in her particular field or subfield. The drive to acquire such methods, however, will not be sufficient. The subject must show proficiency in using such methods by reliably attaining the goals of her motivation. Zagzebski explains that if a virtuous agent is

> truly open-minded, she must actually be receptive to new ideas, examining them in an even-handed way and not ruling them out because they are not her own . . . Similarly if she is intellectually courageous she must, in actual fact, refrain from operating from an assumption that the views of others are more likely to be true than her own.[10]

The manifestation of virtue will vary across contexts. Zagzebski explains that a logician, for instance, will display thoroughness and insight through his advanced deductive and inductive reasoning capacity, while a journalist will display such virtues through perceptual acuity or fact-finding skills. A doctor might display these qualities by carefully adapting lessons from his clinical experience and schooling to the unique circumstances of individual patients in order to yield accurate diagnoses. Possessing virtue, then, involves a steady motivation and a firm understanding of effective methods of belief-formation.

As suggested above, an expert not only adopts effective means of belief formation, and subsequently possesses a deep, unified understanding of her field, but also displays a distinctive method of deliberation. Again, responsibilist theories of virtue seem to lead us in the right direction. According to Hookway, the exercise

[8]I am not here defending a virtue-theoretic analysis of knowledge, but rather suggesting that contextualists broaden their view to incorporate virtues as significant to knowledge ascriptions. This is because I accept the basic contextualist thesis that knowledge standards will vary and it is not clear that virtues will be relevant to identifying knowers in *all* contexts. For a different view, see Christopher Hookway, "How to be a Virtue Epistemologist," in *Intellectual Virtue: Perspectives from Ethics and Epistemology,* eds. Michael DePaul and Linda Zagzebski. (New York: Oxford University Press, 2003): 183–202.) Hookway suggests that the central focus of epistemology move towards an examination of the deliberative process, rather than the analysis of the static cognitive states of justified belief and knowledge. He discusses virtues not as an element of knowledge, but rather as relevant to epistemic evaluation outside of this focus.

[9]Zagzebski, Linda. *Virtues of the Mind.* (Cambridge: Cambridge University Press, 1996).

[10]Zagzebski, 177.

of epistemic virtue consists not in a sort of monitoring and constant awareness of one's deliberative methods, but rather in following the natural course—the questions and issues that occur to her—of one's reasoning. How does such apparently passive reception qualify as the manifestation of virtue and meet norms of good deliberation? Hookway maintains that well-entrenched cognitive character traits established through sufficient training and experience guide one's process of inquiry so that the agent may proceed in a fluid and relatively automatic manner. What this means is that the intellectually virtuous agent will trust the direction of her deliberation without stopping to reflect at every juncture. He maintains that virtuous deliberation

> cannot be a matter of mastering rules which are consciously applied in planning and evaluating deliberations . . . .it is manifested in the fact that distinctive thoughts and questions do *not* occur to you in the course of your deliberation . . . [One learns] not to find certain kinds of considerations salient. Until he has acquired this negative deliberative capacity . . . he cannot perform actions or carry out inquires . . . because he lacks the capacities for deliberation which are required for the successful exercise of those virtues.[11]

An analogy from ethics is instructive. Consider what it means to be benevolent. The benevolent agent does not agonize over whether every situation she encounters requires benevolence, or attempt to develop specific criteria which allows her to identify benevolence-requiring circumstances. Indeed doing so might reflect efforts or attempts to be benevolent, but seems to fall short of actually possessing the trait. The possibility of benevolent action occurs to the benevolent agent in appropriate and sufficient cases, and – this is key—"that this occurs should not be something which [she] consciously [monitors] and [controls]".[12] In the same way, an individual who possesses the cognitive virtue of being observant does not incessantly scan the room for assorted bits of information, but is rather "open to her surroundings, taking notice of things that are interesting and important".[13] The observant agent is discriminately sensitive to certain features of her environment, but without reflecting upon the conditions of her sensitivity. This is not to say that the agent cannot or would not pause, if necessary, to review her methods, say, because she has encountered unexpected results, but that doing so will not be necessary for most judgments. Virtuous deliberation, then, involves

> carefulness [which] is manifested in the fact that he knows when to check inferences and observations and rarely makes mistakes. And his intellectual perseverance is shown in, for example, his ability to acknowledge the consequences of his views without wavering. Such virtues regulate the ways in which we carry out such activities as inquiry and deliberation.[14]

[11]Christopher Hookway, "Epistemic Norms and Theoretical Deliberation" *Ratio* 22 (December 1999): 380–397. Hookway also discusses negative norms of deliberation in his "Epistemic Akrasia and Epistemic Virtue" in Virtue Epistemology: Essays on Epistemic Virtue and Responsibility, eds. Linda Zagzebski and Abrol Fairweather (Oxford: Oxford University Press, 2001): 178–199.

[12]Hookway, "Epistemic Norms" 386.

[13]Ibid., 392.

[14]Hookway, "How to be a Virtue Epistemologist" 187.

The exercise of virtue, according to responsibilists involves both a strong grasp of discipline-specific principles of belief formation and deliberative activity that is appropriately guided by such understanding. This helps explain our epistemic evaluation of the tech and the doctor in PTSD. We presume the doctor's belief, unlike the tech's, is produced and guided by certain intellectual virtues, which he will have acquired throughout his medical training and clinical experience. The doctor will be trained to be attentive to critical features of the patient's situation: Joe's history and demeanor, the effects of other medications, family history, etc, and carefully sift through this information to generate the appropriate diagnosis, thereby portraying the sorts of virtues endorsed by Hookway and Zagzebski: open-mindedness to alternative hypotheses, thoroughness, and insight. It is because we expect Dr. Franklin to display such qualities that we ascribe knowledge to him and not the tech.

What, then, does this mean for the contextualist? In order to account for our intuitions in such cases, contextualists cannot limit epistemic evaluation, and in particular, knowledge attributions and denials, to the subject's ability to respond appropriately to $\sim p$ worlds. Rather, they must allow that in cases where individuals seek expert advice, a subject will qualify as knowing the relevant proposition only by exercising the sorts of virtues discussed by responsibilists. This is because our ascriptions of knowledge in such contexts seem to hinge on the subject's broad understanding of underlying principles in her field and her ability to effectively apply such understanding. I will conclude by considering some plausible worries against my proposal.

One might argue that my argument is subject to a dilemma: if such virtues do not improve the subject's evidentiary relation to *p,* or her ability to track *p* in relevantly similar situations, then considerations of virtue are irrelevant to knowing that *p*. On the other hand, if virtues do improve the subject's epistemic position in just these ways, the differences in epistemic status can be accommodated within the traditionalist contextualist picture. This worry, however, presupposes that the only thing relevant to epistemic evaluation is the subject's performance in contextually relevant counter-factual situations. But this is precisely what is challenged by our intuitions in cases like PTSD. These cases suggest that we are inclined to consider other factors, given that the agent's ability to rule out alternatives and track the truth of *p* does not distinguish her as a knower. What I've argued is that responsibilist virtues offer a good explanation of what these such factors might be.

Another worry is that incorporating a reliabilist form of justification, which focuses on methods of belief formation, might do a better job of accounting for our judgments than the route I've suggested. A reliabilist version of contextualism would claim that in expert scenarios, knowers will have to use extremely reliable methods of belief formation. This makes it the case that the doctor knows that Joe is suffering from PTSD because his methods are sufficiently truth-conducive, while the tech's are not. However, it seems that the very same problem occurs for a form of contextualism. We can imagine a context in which two subjects have used methods of belief formation that are *equally* reliable, but we are nevertheless inclined to reserve our knowledge attributions for only one. Suppose that the tech

has assured you, the patient, that her check-off list allows her to be right 95 % of the time. It still appears prudent for you to reject the tech's claim and return to speak with the doctor. Again, this will not be true in low-standard cases where we are considering single propositions, unrelated to a broad network of facts. Clearly, a bank customer inquiring about whether the bank will be open on Saturdays will not be interested in whether or not the teller has exercised such virtues of fair-mindedness and insight. In expert contexts, we are unlikely to attribute knowledge to a reliable tech because we value the sort of understanding and competence displayed by the doctor. In these contexts, it is not merely the truth-conduciveness of the methods involved, but also whether the cognizer has tapped into the right sorts of facts and integrated them in appropriate ways.

These considerations suggest yet a third alternative: coherentism. According to coherentist theories of justification, a belief is justified in virtue of its membership to a coherent system of beliefs, where coherence depends on logical and probabilistic consistency as well as explanatory and other inferential relations among beliefs in the set. The system of beliefs, then, is the primary unit of justification, and may include either the subject's entire corpus of beliefs, or, more plausibly, some smaller group within this set. Whatever the strengths or weaknesses of coherentism generally, it might make sense to allow that justification in this context where broader understanding is relevant, involves coherence. This allows the contextualist to claim that the doctor knows in PTSD because he possesses a sufficiently interconnected subsystem of beliefs about human psychology and mental illness, but the tech did not know because her beliefs did not bear such relationships. The advantage of using coherentism is that it emphasizes the sorts of links that we value in such contexts.

This move, however, will not help the contextualist for two reasons. First, the information that the doctor organizes and threads together needn't advert to his belief set. Perhaps upon being presented with his patient, the doctor recalls a study in which the effectiveness of a particular PTSD drug was tested, and remembers being somewhat persuaded by the results of this experiment. Note that what the doctor has called into mind, and therefore connected to the patient's diagnosis, is not an existing belief, but rather a slight *inclination* he had toward the researcher's claim, which might then prompt him to consider the treatment, investigate the drug further or ratiocinate on what the results of that study reveal about the nature of the illness. The types of correlations and links valued here are not then confined to coherence among beliefs. While one might expand the coherentist theory to include other mental states, it is not clear that such a move would be uncontroversial or defensible within a coherentist framework.[15] More importantly, even with this modification, coherentist justification cannot accommodate the skill in question, as demonstrated by the next point.

[15]Indeed, Bonjour attempts to include the input of observation states, but maintains that they do not provide justification for resulting perceptual beliefs. Laurence BonJour, *The Structure of Empirical Knowledge* (Cambridge: Harvard University Press, 1985), chapter 6.

Second, the associations made by the doctor might not in fact be supported by his existing system of beliefs. Suppose a new hypothesis strikes the doctor upon examining Joe. This hunch is certainly not entailed by his existing beliefs, and may even be problematic or conflict with some of his current beliefs. This is not to say that the hypothesis is entirely unfounded, but simply that it does not feature the type of supportive relationships that constitute a coherent system. Indeed, there is empirical evidence to suggest that this is the type of cognitive processing employed by successful experts. Pritchard, for example, presents a case in which medical practitioners deviate from the instruction received and beliefs acquired during schooling and residency to treat epilepsy.[16] This cannot be understood in terms of coherence, as it involves the subject's ability to make conjectures that do not cohere well with his existing system of beliefs. Thus, a coherence model does not properly capture the expert's epistemic processes. On the other hand, his open-mindedness about alternative approaches, courage to defend unconventional methods and the honesty to recognize flaws in a previous system all speak in favor of a virtue-contextualist approach to such cases.

13.4 Conclusion

Insofar as traditional contextualism focuses myopically on a subject's ability to eliminate relevant alternatives or track the truth of her belief in close possible worlds, it will fail to account for our judgments in expert contexts, which consequently diminishes intuitive support for the theory. Contextualists, then, can explain our epistemic evaluations in cases of expert inquiry, and thereby reclaim support from ordinary uses of 'knows', by modifying their notion of epistemic strength to include the exercise of virtue.

References

BonJour, L.: The Structure of Empirical Knowledge. Harvard University Press, Cambridge (1985)
Cohen, S.: Contextualism, skepticism and the structure of reasons. In: Tomberlin, J.E. (ed.) Philosophical Perspectives 13: Epistemology, pp. 57–89. Ridgeway, Atascadero (1999)
DeRose, K.: Contextualism and knowledge attributions. Philos. Phenomenol. Res. **52**(4), 913–929 (1992)
DeRose, K.: Solving the skeptical problem. In: Keith, D.R., Ted, W. (eds.) Skepticism: a contemporary reader, p. 206. Oxford University Press, Oxford (1999)
Dretske, F.: The pragmatic dimensions of knowledge. Philos. Stud. **40**(3), 363–378 (1981)
Hookway, C.: Epistemic norms and theoretical deliberation. Ratio **22**, 380–397 (1999)

[16]Pritchard, "Virtue Epistemology and the Acquisition of Knowledge".

Hookway, C.: Epistemic Akrasia and epistemic virtue. In: Zagzebski, L. (ed.) Virtue Epistemology: Essays on Epistemic Virtue and Responsibility. Oxford University Press, Oxford/New York (2001)

Hookway, C.: How to be a virtue epistemologist. In: DePaul, M., Zagzebski, L. (eds.) Intellectual Virtue: Perspectives from Ethics and Epistemology, pp. 183–202. Oxford University Press, New York (2003)

Lewis, D.: Elusive knowledge. In: DeRose, K., Warfield, T. (eds.) Skepticism: A Contemporary Reader, pp. 220–238. Oxford University Press, Oxford (1999)

McGrath, M.: Evidence, pragmatics and justification. Philos. Rev. **111**(1), 67–94 (2002)

Pritchard, D.: Virtue epistemology and the acquisition of knowledge. Philos. Explor. **8**(3), 229–243 (2005)

Zagzebski, L.: Virtues of the Mind. Cambridge University Press, Cambridge (1996)

Chapter 14
Defeasible Argumentation in African Oral Traditions. A Special Case of Dealing with the Non-monotonic Inference in a Dialogical Framework

Gildas Nzokou

Abstract The main claim of **the present** paper is to defend that some specific oral **debate** forms of the African traditions seem to correspond structurally speaking to non-monotonic reasoning in a way that is not that different from nowadays argumentation-based approaches of legal reasoning within the context of western juridical systems. So, the aim of this survey consists in two points: on the one hand, we will show that polemical debates in African oral traditions implement systematically a non-monotonic inference, that is closed to what Aristotle termed by "dialectical arguments"; on the other hand, we are suggesting a way to **deal** with non-monotonic inference in a dialogical framework.

Keywords Argumentation • Inference • Non-monotony • Dialogical logic • Oral traditions • Proverbs

14.1 Introduction

The exposition plan of the present survey follows from a methodological motivation. Indeed, a study about argumentation procedures in African oral traditions relates to both, anthropological issues and logical technicalities. Accordingly, we shall first talk about the general cultural futures of our topic and second, we will develop the technical aspects.

In the following paragraphs I start with a brief description of the type of arguments in consideration – those arguments that make use of proverbs – and then I will link them with their cultural function and role.

G. Nzokou (✉)
Centre d'Études et de Recherches Philosophiques, Faculté des Lettres et Sciences Humaines, Université Omar Bongo, Libreville, Gabon
e-mail: nzokou_gildas@yahoo.fr

J. Redmond et al. (eds.), *Epistemology, Knowledge and the Impact of Interaction*, Logic, Epistemology, and the Unity of Science 38, DOI 10.1007/978-3-319-26506-3_14

In African oral traditions, argumentative practices paradigmatically include the use **of** certain kind of strategic premises, namely the proverbial ones, in order of increasing the strength of the argument that is being developed. This device enables to infer reasonably some conclusions. However, at a next step of the debate, the conclusion so far achieved, might be withdrawn by the addition of new information.

Moreover, when an epistemic agent supports a thesis, he makes use of a set of ordinary premises – these express a factual content – plus a strategic one, the proverbial sentence. The point of the argument is to show that the thesis can be justified by making it apparent that it follows as the conclusion of the given factual premises and the proverbial sentence. From a cultural point of view, no inference can be done without the **rationality** weight of a proverb. This feature of the proverb allows **the crossing** from premises to the conclusion, because it works both as a primitive proposition and as a special inference rule (into the epistemological frame of the traditions considered here).

More generally, in the context of the African oral tradition proverbs constitute fundamental bricks of knowledge and knowledge principles. In fact, proverbs encode a **synthesis** of different types of phenomena (political, biological, sociological, spiritual, etc.) that the tradition raises to a paradigmatic standard of knowledge principle. Thus, in argumentation contexts, the occurrence of proverb within the premises corpus warrants the rational and gnoseological legitimacy of the inferred conclusion.

There is a very important and essential moment in the use of a proverb. It is the hermeneutic phase which consists in establishing an analogy between the generic image which is the proverb and the factual situation which is at stake. If this interpretative phase is led suitably, the use of proverb appears as relevant. This idea will be formally taken into account by making use of an analogy relation introduced at the object language level. So, structurally, the occurrence of a proverbial sentence, as a strategic premise, allows crossing from the premises to the conclusion, *via* the analogy that one develops just before. Our formalization does not describe the process of the analogy, since this requires a far more rich structure than assumed in the present framework.[1] In fact, we will describe the argumentation process after the analogy has been settled. However we should always keep in mind that the reasoning process that takes place in argumentation with proverbs involves both the inferential and the analogical moves.

From a methodological point of view, we need a suitable inference relation and a suitable theoretical framework to take relevantly account **of** an abstract reformulation of such an argumentation form. And, for this purpose, the dialogical approach to logic seems perfectly suitable.

[1] In recent work, Bernadette Dango (2015) suggested that the establishment of an analogy could be thought as finding some kind of mapping that displays the transition from one epistemic state to a second one, such that the latter is more specific than the first one. Moreover, according to Dango the didactic role of proverbs is to train in creating such kind of mappings. In other words, according to Dango; the way to teach and transmit knowledge gathered by a given tradition amounts teaching on how to apply it in a specific situation.

In fact, Dialogical logic (DL) seeks to recover both, the philosophical and technical link between argumentation and logic (logic as Agon) *via* the development of pragmatist semantics. This semantics provides the basis for the notion of formal strategy by the means of which inference is understood dynamically, i.e., as a kind of a rational interaction of agents.

The historical context and its actual developments had been discussed at length by Rahman and his teams of Saarbrücken and Lille.[2] So, I will concentrate on the rules of the system.

I would just **remind** that the dialogical approach was inspired by Wittgenstein's meaning as use, that is, the meaning has to be drawn from the context of use. The basic idea of the dialogical approach to logic is that the meaning of the logical constants is given by the norms or rules for their use – in fact, dialogical logic is the first theoretical framework that implements Wittgenstein's idea of meaning as use in logic. This feature of its underlying semantics quite often motivated the dialogical approach to be understood as a pragmatist semantics.[3] The point is that those rules that fix meaning may be of more than one type, and that they determine the kind of reconstruction of an argumentative and/or linguistic practice that a certain kind of language games called dialogues provide.

However, given that the arguments constructed by means of proverbs (as their strategic premise) are defeasible because of the inherent possibility that a counter-proverb can be used by a challenger, and so succeed to block the job of a precedent proverb, it naturally seems that we are facing a revision phenomenon on the sets of premises. This is why we use also a small fragment of belief revision to formalize the structural work of proverbs within the argument's corpus.

Let's begin by giving a sketch of the technical tools we will use in the present survey.

14.2 Epistemic Dynamic by Means of Belief Revision

14.2.1 Epistemic States and Belief Sets

Usually, in the context of belief-change, one considers the epistemic states in the model of belief sets closed under the classical consequence. This assumption of deductive closure is expressed by the following principle of reflexivity:

- $K = Cn(K)$

[2]See Rahman 1993, Rahman and Rückert 2001, Rahman and Keiff 2004, Rahman et al. 2009, Rahman and Tulenheimo 2009.

[3]Quite often it has been said that dialogical logics has a pragmatic approach to meaning. I concede that the terminology might be misleading and induce one to think that the theory of meaning involved in dialogic is not semantics at all. Helge Rückert proposes the more appropriate formulation pragmatistische Semantik (pragmatist semantics).

That means that the sets of beliefs for the epistemic agents contain all their proper consequences (possibly infinite). And these closure operations under the classical consequence possess the following three important properties:

- Inclusion (or reflexivity): $K \subseteq Cn(K)$
- Idempotence: $Cn(A) = Cn(Cn(A))$
- Monotony: if $A \subseteq B$ then $Cn(A) \subseteq Cn(B)$.

It's that monotonicity which is dropped anyway within the development of the defeasible argumentation form. It means that the sets of premises will not be closed under the classical deductibility, so the consequence relation will appear as being unstable – or to put positively, the consequence relation will be open to changes.

Now, let us briefly present the part of the theory change we are interested in for the requirements of this survey. It's precisely about revision and contraction operations.

14.2.2 Epistemic Dynamic Modes: Contraction and Revision

We work here with the AGM model. In such a model, the rationality's criteria for beliefs change are summarized by the following axiomatic system:

(A) **Contraction**.

- (K^- 1) let K be any set of beliefs or knowledge, and A any single belief content, the contraction of A from K, written as $K \div A$, is also an epistemic set.
- (K^- 2) $K \div A \subseteq K$.
- (K^- 3) If $A \notin K$, then $K \div A = .K$.
- (K^- 4) If $\nvdash A$, Then $A \notin K \div A$.

The remaining axioms are the following:

- (K^- 5) : If $A \in K$, then $K \subseteq (K \div A)^{+}_{A}$.
- (K^- 6) : If $\vdash A \leftrightarrow B$ then, $K \div A = K \div B$.
- (K^- 7) : $K \div A \cap K \div B \subseteq K \div (A \wedge B)$
- (K^- 8) : If $A \notin K \div (A \wedge B)$, then $K \div (A \wedge B) \subseteq K \div A$.

(B) **Revision**:

The revision happens when an epistemic input A, accordingly to an epistemic basis K, contradicts some element already contained in K. That is, the arrival of A in K entails an absurdity. Now, given that the new information (i.e. the input) has the benefit of priority on the ancient one, for avoiding the inconsistency in the epistemic system, the methodological principles require that one first withdraws the element of K which is contradicting A; next, one can add A in K previously **cut by** ¬A. That is resumed in the Levy's identity we shall discuss below. Let us say that, the revision responds to the need of maintaining a rational equilibrium in belief system; stated otherwise, one needs to get a consistent epistemic agent rather than an absurd one.

Now, an interesting point concerning the revision operation is that there is no monotonicity yet; even though K may be included in H, that does not imply systematically that K*A is included in H*A.

This theoretical setting shows that the modelling of epistemic dynamics appears as fundamental within a survey concerning the understanding of human reasoning. Therefore, the link with the argumentation forms studied here is obvious. But, let us go **for** the essential:

For any set of beliefs K and single belief A,

- (K* 1) K^*_A is also a belief set.

The second postulate – which is a success postulate – states that the input is a belief accepted by K

- (K* 2) $A \in K^*_A$.
- (K*3) $K^*_A \subseteq K^+_A$
- (K*4) If $\neg A \notin K$, then $K^*_A \subseteq K^+_A$
- (K*5) $K^*_A = K_\perp$ iff $\vdash \neg A$.

The rest of the axiomatic system is not immediately relevant for our actual purposes. But, before continuing the exposition, let us see an interesting example of non-monotonicity caused by the revision operation on two propositional sets having the same basis.

Let K_1 and K_2 be some sets of beliefs such that: $K_1 = \{p, q\}$ and $K_2 = \{p, p \leftrightarrow q\}$. From K_1 one can infer $p \leftrightarrow q$, so $(p \leftrightarrow q) \in Cn(K_1)$. Next, one considers K_1 and K_2 as the same since they have identical consequences. Now, if one applies a uniform revision on both belief sets, one sees how the monotony disappears as follows:

$(K_1)^* \neg p = \{p, q\}^* \neg p = (\{p, q\} \div p)^+ \neg p$, what amounts to $(K_1)^* \neg p = \{\neg p, q\}$ (let us term this K_1'). At the same time:

$(K_2)^* \neg p = \{p, (p \leftrightarrow q)\}^* \neg p = [\{p, (p \leftrightarrow q)\} \div p]^+ \neg p$ what amounts to $(K_2)^* \neg p = \{\neg p, p \leftrightarrow q\}$ (let us call this K_2').

Next one must consider $Cn(K_1') = \{\neg p, q, \neg p \leftrightarrow q\}$, i.e. $((\neg p \rightarrow q) \wedge (q \rightarrow \neg p)) \in Cn(K_1')$, meanwhile $Cn(K_2') = \{\neg p, p \leftrightarrow q\}$, i.e. $((p \rightarrow q) \wedge (q \rightarrow p)) \in Cn(K_2')$. From there, we can see clearly how the revision operation induces the non monotonicity of the inference.

14.2.3 Proverbial Sets as Beliefs Bases

In reaction to what precedes above, it appears natural to consider the stocks of proverbs and the other sentences used when setting out arguments on the model of belief bases.

The central idea here is that it seems more natural to think of epistemic states as some kind of structures which can be modelled by means of finite sets of propositional contents. These latter are being knowledge and beliefs explicitly put

into the cognitive frame of the epistemic agent. Stated otherwise, a belief base is a finite set of beliefs and knowledge which the agent is clearly aware of. So the problem of the logical omniscience (entailed by the classical closure) is avoided. The epistemic change happens only on the elements of these belief bases. This is how this idea is formally presented:

14.2.3.1 Definition: Belief Base

Any set of proposition is a belief base. Let K and A be two sets of propositions: A is a belief base for K iff:

- $\mathbf{K} = \mathrm{Cn}(\mathbf{A})$

That induces the following:

- α is a belief iff $\alpha \in \mathrm{Cn}(\mathbf{A})$
- α is a basic belief iff $\alpha \in \mathbf{A}$
- α is a simply inferred belief iff $\alpha \in \mathrm{Cn}(\mathbf{A})\backslash\mathbf{A}$[4]

Some Remarks on the Notation in some places in the text, we will write *prem*(A) and *Conc*(A) to indicate the set of A's premises and, respectively, a conclusion for the argument A.

After this brief presentation of the theory's change fragment needed to formalize the non-monotonicity of the inference relation in use in defeasible argumentation, let us now give a brief sketch of Dialogical Logic.

14.3 Dialogical Logic

14.3.1 Generalities on Dialogical

In a dialogue two parties argue about a thesis respecting certain fixed rules. The player that states the thesis is called Proponent (**P**), his adversary, who contests the thesis, is called Opponent (**O**). In its original form, dialogues were designed in such a way that each of the plays end after a finite number of moves with one player winning, while the other loses. Actions or moves in a dialogue are often understood as utterances[5] or as speech-acts.[6]

[4]Presentation based on Hansson, where this notation is precised by saying that, for any two sets X and Y, X\Y is the set of elements of X which are not in Y.

[5]*C*f. Rahman and Rückert 2001, 111 and Rückert 2001, chapter 1.2.

[6]Cf. Keiff 2007.

Some points of precision are essential for a good comprehension of the dialogical approach:

1. We have to distinguish between local (rules for logical constants) and global meaning (included in the structural rules) – it is not the difference between introduction-elimination rules and structural rules.
2. The player independence of local meaning (recall that in a tableaux system, for example, the meaning of the logical constants is dependent on the sides: True-rules and False-Rules)
3. The distinction between the play level (local winning or winning of a play) and the strategic level (global winning; or existence of a winning strategy) – the notion of play does not correspond neither to winning in a model nor to a branch in a proof-tree)
4. The notion of formal play and strategy (this does not correspond to true in any model, but true in a play where the proponent does not know about the justification of the atomic formulae)

14.3.2 Dialogical Logic in proper

Despite the existence of many different formulations, we present the system by following here the notations and definitions given by Nicolas Clerbout in his recent book on dialogical semantic and meta-dialogical.[7] However, for the sake of **clarity** and detail, we **will** use the standard exposition.

Let the language $\mathcal{L}$ be composed of the standard components of first order logic with connectives $\wedge$, $\vee$, $\rightarrow$, $\neg$, and two quantifiers $\forall$, $\exists$ – the conjunction might be indexed yielding $\wedge_{i,}$ where $i \in \{1, 2\}$, such that "$\wedge_1$" stands for the first conjunct from left to right,

- Small letters (p, q, ...) for atomic formulæ,
- Greek letters (α, β, ...) for formulæ that might be complex,
- Capital italic bold letters (**A**, **B**, **C**, ...) for predicators,
- Constants k_i, where $i \in N$, and
- Variables x, y, z, ...

We will also need the following special symbols and idioms:

Force symbols: ? and !. The **question** mark might be combined by means of specific rules yielding: $?\wedge_i$, $?\vee$, $?_{ki}$, $?_\exists$. The sign "?" ("!") signalises that a given move is a challenge (defence).

[7] Clerbout (Nicolas), 2014; La sémantique dialogique. Notions fondamentales et éléments de métathéorie, London, College Publications, „Cahiers de Logique et d'épistémologie" series, Vol. 21.

Rank-idioms[8]: **r** := N, **r′** : = N' where N and N' are natural numbers and "**r**" ("**r′**") stand for "repetition rank". Accordingly, "**r** = 1" and "**r′** =2" signalise that the repetition ranks chosen are 1 and 2.

Def. 1: An expression of **L** is either a term, a formula, a force symbol or a rank idiom.

Def. 2: Every expression ***e*** of our language can be augmented with labels **P** or **O** (written **P**-***e*** or **O**-***e***, called (dialogically) signed expressions), meaning in a game that the expression has been uttered by **P** or **O** (respectively). We use X and Y as variables for **P**, **O**, always assuming X ≠ Y.

Def. 3: A move μ is dialogically signed expression X-***e***.

14.3.2.1 Local Meaning

Particle Rules

In dialogical logic, the particle rules are said to state the local semantics: what is at stake is only the request and the answer corresponding to the utterance of a given logical constant, rather than the whole context where the logical constant is embedded.

The following table displays the particle rules, where X and Y stand for any of the players **O** or **P**:

∨, ∧, →, ¬, ∀, ∃,	Challenge	Defence
X: α∨β	**Y**: ?-∨	**X**: α
		or
		X: β
		(**X** chooses)
X: α∧β	**Y**: ? ∧1	**X**: α
	or	respectively
	Y: ? ∧2	**X**: β
	(**Y** chooses i ∈ {1, 2)	
X: α → β	**Y**: α	**X**: B
	(**Y** challenges by uttering α and requesting B)	
X: ¬α	**Y**: α	—
		(no defence available)
X: ∀xα	**Y**: ?k	**X**: α [x/k]
	(**Y** chooses)	
X: ∃xα	**Y** ? ∃	**X**: α [x/k]
		(**X** chooses)

[8]In the present survey we don't use the rank-device which is very relevant and useful for the new reformulation and new perspectives on this topic.

In the diagram, α[x/k] stands for the result of substituting the constant k for every occurrence of the variable x in the formula α.

One interesting way to look at the local meaning is to see it as rendering an abstract view (on the semantics of the logical constant) that **makes a distinction** between the following types of actions:

(a) Choice of declarative utterances (=:disjunction and conjunction).
(b) Choice of interrogative utterances involving individual constants (=: quantifiers).
(c) Switch of the roles of defender and challenger (=: conditional and negation). As we will discuss later on we might draw a distinction between the switches involved in the local meaning of negation and the conditional).

Let us briefly mention two crucial issues

14.3.2.2 Plays and Games

Def. 4
A play is a legal sequence of moves that complies with the moves of the particle rules described above and the structural rules to be described below.

Def. 5
The dialogical game for α (**D**α) is the set of all plays with thesis α in the sense of the starting rule SR-0 (given below).

Def. 6
An X-terminal play for α is a play △ in **D**α such that the last member of △ is a X-move and there is no Y-move μ such that $\triangle' = \triangle \frown \mu$ is a play in **D**α – where "$\triangle \frown \mu$" stands for a play that extends △ with the move μ. A dialogue for α is an X-(or Y)-terminal play

Def. 7
For any sequence Σ of moves μ, the function π^{Σ} assigns to each member of Σ a (non-negative) position (-number): if μ is the *i*-th member of Σ, then $\pi^{\Sigma}(\mu) = i - 1$. Thus, if μ is the first member of the sequence the function will assign to this move the position 0. (If there is no ambiguity on which the sequence is involved we will write simply $\pi^{\Sigma.})$.

Def. 8
For any sequence Σ of moves μ, the partial function $\mathcal{J}^{\Sigma}$ assigns to each member of Σ, that is not a rank idiom (**r**=:N, $\mathbf{r}' = \mathrm{N}'$) and that has a position bigger than 0 a pair [μ′, Z] such that Z ∈ {?, !}, μ′ is a move of the antagonist player and $\pi^{\Sigma}(\mu') < \pi^{\Sigma}(\mu)$.

The intended interpretation of $\mathcal{J}^{\Sigma}(\mu)$ is that each move μ of the sequence that is neither the thesis (that has position 0) nor is a rank idiom either challenges a previous move μ' or is a defence against a previous challenge μ', where μ' is a move of the antagonist player.

14.3.2.3 Global Meaning

Structural Rules

(SR-0) (Starting rule):

Every play in the dialogical game for α (**D**α) starts with a move of **P** uttering α such that its position is 0. It provides the topic of the argumentation and is called the thesis of the play.

Moves are alternately uttered by **P** and **O**. That is, given a play $\triangle$ in **D**α and a move μ in $\triangle$, it is the case that if $\pi(\mu)$ is even then it is a **P**-move. Dually moves with odd positions are **O**-moves.

Comment: The proviso if possible relates to the utterance of atomic formulae. See formal rule (SR 2) below.

(SR-1) (no delaying tactics rule):

Both **P** and **O** may only make moves that change the situation.

After the move that sets the thesis players **O** and **P** each choose a natural number **r** and **r**′ respectively (termed their repetition ranks).

In the course of the dialogue, **O** (**P**) may attack or defend any single (token of an) utterance at most **r** (or **r**′) times.

Notice that the repetition rank does apply neither to the move that fixes a repetition rank nor to the utterance of the thesis: it fixes only the number of utterances of a move that is a challenge or a defence.

Thus, each move whose position is bigger than 2 is either a challenge or a defence (see Def-8) – since in position 0 the thesis is uttered and positions 1 and 2 are occupied by moves that result from the choice of a repetition rank.

(SR-2) (Formal rule):

P may not utter an atomic formula unless **O** uttered it first.

More precisely, a sequence $\triangle$ in the dialogical game **D**α,where the thesis α is a complex formula, constitutes a formal play for α if for any atomic formula q in $\triangle$ it is the case that

If $\mu = \mathbf{P}\text{-q} \in \triangle$, then there is a $\mu' = \mathbf{O}\text{-q} \in \triangle$ such that $\pi^{\triangle}(\mu') < \pi^{\triangle}(\mu)$.

Atomic formulae cannot be challenged (i.e., for any atomic formula q occurring in a play $\triangle$ there is no move in that play of the form [q, ?])

The dialogical framework is flexible enough to define the so-called material dialogues that assume that atomic formulae have a fixed truth-value. That is very important for our purpose, for, the argumentation form we try to reconstruct here is about a kind of material dialogue, where the cultural background looks like an oracle providing the needed argumentative resources (i.e. proverbs) to each of the antagonists.

(SR *2) (Rule for material dialogues):

Only atomic formulae standing for true propositions may be uttered. Atomic formulae standing for false propositions cannot be uttered.

(SR 3) (Winning rule):

X wins iff it is Y's turn but he cannot move (either challenge or defend).
More precisely, X wins a play $\triangle$ for α in $\mathbf{D}\alpha$ iff it is X-terminal (see Def-6)

Global Meaning
These rules determine the meaning of a formula where a particle occurs as a main operator in every possible play.

(SR 4c) (Classical rule):

In any move, player X (Y) may challenge a complex formula uttered by his partner or he may defend himself against any challenge (including those challenges that have already been defended once) at most **r**-times ($\mathbf{r}'$-times).

or

(SR 4i) (Intuitionist rule)[9]:

In any move, player X (Y) may challenge a (complex) formula uttered by his partner at most **r**-times ($\mathbf{r}'$-times) – where **r** ($\mathbf{r}'$) is the correspondent repetition rank – or he may defend himself against the last challenge that has not yet been defended – the latter condition on defences has priority over **r.** (see example 1).

In the dialogical approach validity is defined *via* the notion of winning strategy. Informally, a winning strategy for X means that for any choice of moves by Y, X has at least one possible move at his disposal such that he (X) wins:

Validity (Definition)
A formula is valid in a certain dialogical system iff **P** has a formal winning strategy for this formula. To be more precise:

[9]In the standard literature on dialogues, there is an asymmetric version of the intuitionist rule, called E-rule since Felscher (1985).

- α is classically valid if there is a winning strategy for **P** in the dialogical game $\mathbf{D_c}\alpha$
- α is intuitionistically valid if there is a winning strategy for **P** in the dialogical game $\mathbf{D_i}\alpha$.

Before we tackle the notion of strategy more thoroughly let us point out some features of the dialogical notion of validity and display then some examples:

Examples
In the following examples, the outer columns indicate the numerical label of the move, the inner columns state the number of a move targeted by an attack. Expressions are not listed following the order of the moves, but writing the defence on the same line as the corresponding attack, thus showing when a round is closed. Recall, from the particle rules, that the sign "—" signalises that there is no defence against the attack on a negation.

For the sake of simplicity we will assume the following rank choices:

O- $\mathbf{r} := 1$
P-$\mathbf{r}' := 2$

Ex. 1: Classical and Intuitionistic Rules
In the following dialogue played with classical structural rules **P**' move 4 answers **O**'s challenge in move 1, since **P**, according to the classical rule, is allowed to defend (once more) himself from the challenge in move 1. **P** states his defence in move 4 though, actually **O** did not repeat his challenge – we **signal** this fact by inscribing the not repeated challenge between square brackets.

O			P		
				$p \vee \neg p$	0
1	$?_\vee$	0		$\neg p$	2
3	p	2		—	
[1]	$[?_\vee]$	[0]		p	4

Classical Rules. P Wins
In the dialogue displayed below about the same thesis as before, **O** wins according to the intuitionistic structural rules because, after the challenger's last attack in move 3, **t**he intuitionist structural rule forbids **P** to defend himself (once more) from the challenge in move 1.

O			P		
				$p \vee \neg p$	0
1	$?_\vee$	0		$\neg p$	2
3	p	2		—	

Intuitionist Rules. O Wins
We have to add now some extra rules which take account of the specificity of the argumentative form we intended here.

Structural Rule S R *2.1 (Introduction of an Extra Premise)
The Opponent is the only one allowed to introduce an extra premise[10] which is not a proverb, but this extra premise must ever be joined to a proverb. Each of the players, X and Y, is allowed to introduce a proverbial sentence at any step of the dialogue.

S R *3.1: Winning Rule for Defeasible arguments:
The antagonist X (Y, respectively) who has succeeded to maintain his thesis, after an assumed maximal and legal set of moves, wins the dialogue. In other words, the last move defeating the adversary's argument leads to winning the dialogue.

Definition: Worthiness of a Defeasible Argument
A defeasible argument is dialogically worthy if and only if the general conditions of the dialogue provide a material winning strategy for the Proponent against all possible moves of the Opponent. Stated otherwise, for any move of the Opponent, the Proponent will always find, through the general conditions of the dialogue, all needed and relevant argumentative resources for defeating the Opponent counter-argument.

S R *4: Challenge Against an Argument
In a dialogical system, the challenge of an argument is made as follows:

(i) One of the players concedes the premises of the antagonist while asking for a justification of his conclusion
(ii) If, it is the Opponent who challenges the argument of the thesis, he may introduce an extra-premise followed by a proverbial sentence. However, he may also state directly a counter-proverb which he will add to the set of premises, imposing by this, a revision on the set of these premises.

[10]The rule concerning the introduction of an ordinary extra-premise [R S*2.1] takes account of the fact that, in a controversy the part which introduces the debate usually utters – at least, that is supposed – all statements useful for sustaining his position, meanwhile omitting one or more details which could weaken this position. The same holds for juridical antagonism, where the litigant is supposed to present, since the start of the debate, all elements of the accusation providing the strength of his complaint. It becomes impossible, during the controversial exchange of respective arguments, that the litigant introduces some new statements. This is the ethic motivation of this formal restriction relative to the introduction of extra-premise under which the proponent plays.

However, intending the ideal conditions of equilibrium of a debate, it is necessary to allow the challenger the possibility **of bringing** the details of precision lacking in the initial statement (complaint) of the litigant, by introducing a unique extra-premise. Furthermore, the impossibility for the proponent to attack this extra-premise is simply due to the fact that one cannot rebut or contest a statement which hasn't been uttered yet by a protagonist, though the existence of this statement is supposed beforehand as one of the material conditions of the dialogue. In fact, we must keep in mind that we are in the context of material dialogues and of non-monotonique argumentation process.

Particle Rules

All the standard particle rules are kept; we only add the one relative to the revision operation. The dialogical local meaning for the analogy relation, which introduced a proverbial sentence in the play, is a special case: this relation is neither a particle rule, nor a structural one. But it is necessary to formally explicit this analogy. For, thanks to it, one can rationally perform an argument using proverbs. So, the local semantic of a proverb is given by the following rule which fixes the conditions of an analogy:

Let $\mathbb{P}$ be a proverb and "$\approx$" an analogy relation. The syntactical form of the proverb is such that $\mathbb{P} = \boldsymbol{\Phi} \rightarrow \boldsymbol{\Psi}$

Assertion $\langle\ !\ \rangle$	Attack $\langle\ ?\ \rangle$	Defence $\langle\ !\ \rangle$
X-!- $\boldsymbol{\Phi} \rightarrow \boldsymbol{\Psi}$	(1): Y-? $\approx \boldsymbol{\Phi}$	$(1')$: X-!- $\boldsymbol{\Phi} \approx (\varphi_1 \vee \ldots \vee \varphi_n)$
		(where φ_i is one of the ordinary premises of the argument)*
	(2): Y-? $\approx \boldsymbol{\Psi}$	$(2')$: X-!-$\boldsymbol{\Psi} \approx \delta_i$ (where δ_i is one of the concessions brought forward during the dialogue and which is the conclusion intended of the argument)

*Remember that the type of arguments at stake here are structurally composed by a set of ordinary premises (statements of facts) plus a strategic one (the proverbial sentence) which enables to reasonably infer a non-fully deductive conclusion. Schematically one has the following configuration:
(a)- ordinary premises or statements of facts: "*the American army and the US air forces are joined to battle in Afghanistan. They are facing an asymmetric war imposed by Taliban*." "*American are the greatest military forces in the world*"
(b)- Strategic premise (the proverbial sentence): "*However, as said by the ancients,**even a tiger becomes like a little cat when it is out of its lands***". [insinuated interpretation: "even the greatest military forces look like second-rate fighters when facing a non conventional war"]
(c)- Conclusion intended: "*so, the US air forces and Army will not win this asymmetric war against Taliban*".

Now, one can make correspond the syntactical form of any proverb with its counterpart in the natural language. The following is about our present proverb: "**If a tiger is out of its lands**" $= \Phi$, "**then it becomes like a little cat**" $= \Psi$ $[\boldsymbol{\Phi} \rightarrow \boldsymbol{\Psi}]$

Let us now describe the local meaning of the revision operator:

Assertion $\langle\ !\ \rangle$	Attack $\langle\ ?\ \rangle$	Defence $\langle\ !\ \rangle$
X-!- $\Gamma^*\{\delta\}$	Y-?$_*$	X-!- $(\Gamma \div \neg\delta)_\delta{}^+$[11]
X- !- $\Delta^+{}_\delta$	Y-?$_+$	X- !- $\Delta \cup \{\delta\}$

Before adapting our argumentation theory to the dialogical system, let us **point out** the main features of the language used for our purpose.

[11]Here, the revision operation performs the Levy's identity which is constituted of two steps: a) first, contract the database (in the instance, it's the set of all the argument's premises) from the negation of the revision's formula; next, extend this contracted database by adding the revision's formula itself.

14.4 The Abstract Argumentation Framework Grounded in the Proverbial Language $\mathcal{L}_{\mathbb{P}}$. The Non-Deductive Closure of the Language

The set of propositions we make use of, namely the proverbial ones, are not closed under the classical deduction. In fact, because of all counter-intuitive implications of the classical logical consequence (infinity of inferred consequences, implication of the infinite set of tautologies, etc.), we naturally choose to consider the sets of proverbial sentences as belief basis. The inference used here will be non-fully deductive. The reason is that one always can find a counter-proverb against another one; that explains the defeasibility of arguments using proverbial sentences as their strategic premises.

Indeed, like default rules, proverbial sentences, taken as special rules of inference, are really non-fully deductive. The genesis of these propositions is such that these are yielded by (empirical) inductive generalization. For this, the conclusions they allow to infer are unstable. So we use the notation "$\Rightarrow$" rather than a usual turnstile for representing this inference relation which is not strictly deductive.

Now, let us see in concrete terms, how this is adapted to the dialogical framework.

1st Possibility Challenge by concession of the thesis premises.

O			**P**		
			$(\varphi \wedge \psi), (\mathbf{\Phi} \rightarrow \mathbf{\Delta})^{*} \Rightarrow \delta$		0
1	$?\approx \mathbf{\Phi}$	0	$\mathbf{\Phi} \approx \varphi$		2
3	$?\approx \mathbf{\Delta}$	0	$\mathbf{\Delta} \approx \delta$		4
5	$(\varphi \wedge \psi) \wedge (\mathbf{\Phi} \rightarrow \mathbf{\Delta})$	0	δ	☺	**14**
7	$\varphi \wedge \psi$		5	$? \wedge_1$	6
9	φ		7	$? \wedge_1$	8
11	$\mathbf{\Phi} \rightarrow \mathbf{\Delta}$		5	$? \wedge_2$	10
13	$[\mathbf{\Delta} \approx \delta]$	δ	11	φ $[\mathbf{\Phi} \approx \varphi]$	12

*We keep here the proverbial formula in bold to distinguish it from the other premises

**The proponent wins this basic dialogue by simply showing the actuality of his argument's conclusion after the opponent has accorded the concession of all premises. Here the strategy is based on the interpretative phase, where the opponent asks the proponent for a suitable reading of his proverb (moves 1 and 3 of the opponent). The proponent gives actually a relevant reading of his proper proverb and therefore infers naturally the conclusion of his argument by linking it to certain premises by means of analogy.

Notice that this first possibility of challenge, which is very basic, intends essentially to show how to dialogically deduce the conclusion of an argument using proverbs in the premises corpus.

Some Remarks on the Analogy Relation As already stated, this relation is not really a logical particle; this is why there is no challenge allowed on it. More precisely, the analogy is used to show the relevance of some proverbial sentence in some dialogical, conversational and argumentative contexts. On the other hand, one cannot challenge this relation since, in fact, the dialogue describes what follows once the analogy has been established.

2nd possibility Challenge by Revision on the thesis' premises (the revision formula is a counter-proverb)

O			**P**	
			$(\varphi \wedge \psi), (\mathbf{\Phi} \rightarrow \Delta) \Rightarrow \delta$	0
1	$[(\varphi \wedge \psi), (\mathbf{\Phi} \rightarrow \Delta)]^{*}_{(\mathbf{\Psi} \rightarrow \neg\Delta)} \Rightarrow \neg\, \delta$	0		
3	$[((\varphi \wedge \psi), (\mathbf{\Phi} \rightarrow \Delta)) \div (\mathbf{\Phi} \rightarrow \Delta)]^{+}_{\mathbf{\Psi} \rightarrow \neg\Delta} \Rightarrow \neg\delta$	0	1 $?_{*}$	2
5	$(\varphi \wedge \psi), (\mathbf{\Psi} \rightarrow \neg\Delta) \Rightarrow \neg\delta$		3 $?_{+}$	4
7	$\mathbf{\Psi} \approx \psi$		5 $?\approx \mathbf{\Psi}$	6
9	$\Delta \approx \delta$		5 $?\approx \Delta$	8
			5 $(\varphi \wedge \psi), (\mathbf{\Psi} \rightarrow \neg\Delta))$	10
11	$?\wedge_1$	10	$\varphi_i \wedge \psi$	12
13	$?\wedge_2$	12	ψ	14
15	$?\wedge_2$	10	$\mathbf{\Psi} \rightarrow \neg\Delta$	16
17	$[\mathbf{\Psi} \approx \psi]$ ψ	16	$\neg\delta$ $[\Delta \approx \delta]$	18
19	δ ☺	18		

Explanations
This is a dialogue **won** by the opponent. The winning process consisted of the introduction of a counter proverb which enabled to infer a counter conclusion that means the construction of a counter thesis.

Move 1 the opponent challenges the thesis of the dialogue by using directly a counter proverb, imposing by there a revision operation on the set **of** the argument's premises. In fact, there is an epistemic revision on the informational database which constitutes the set of premises.

Move 2 the proponent counter attacks by asking the opponent for details and precisions about the revision operation imposed by the use of the counter proverb.

Move 3 the opponent gives the precisions about this revision operation at stake, i.e. first, contract the basic set of premises by the negation of the new information – the counter proverb's negation – and next, to extend this contracted database by adding the counter proverb's formula. In other terms, that is an application of Levy's identity we saw previously in the axiomatic for belief revision.

Move 4 it is the continuity of move 2; the proponent asks for the second step of the levy's identity. That is, the extension operation after the contraction one. The answer to this challenge is given in the move 5 by the opponent.

Moves 6 and 8 the proponent's challenge now is to ask for the relevance of each constituent of the counter proverb formula (introduced by the opponent) with some premises and the conclusion of the argument in assessment. What the opponent answers to in moves 7 and 9 (actually the opponent shows a relevant analogy for the counter proverb's antecedent and also for its consequent).

Move 10 the proponent now concedes the new revised database – that is the new configuration on the set of premises – and asks for a justification of the argument's conclusion.

Moves 11, 13, and 15 the opponent challenges the different compound formulae conceded by the proponent in move 10. These challenges are performed by means of the standard particle rules.

Moves 17 the opponent challenges the strategic premise (the counter proverb) conceded by the proponent from move 10 and re-instantiated in move 16. This challenge is done in recalling the analogy established in moves 7 and 9.

Move 18 the proponent answers to this later attack by stating the consequent of the counter proverb formula (which is now one of his proper concessions). This consequent is a negative formula.

Move 19 the opponent challenges the negation from move 18 by stating its dual. This challenge on negative literal cannot be answered according to the standard dialogical particle rules. Therefore, the opponent wins the dialogue.

14.5 Final Remarks About the Ranks

The reader certainly noticed that we didn't use the ranks device. Simply, we didn't state explicitly and formally their use into our dialogical tableaux frame. But, it is easy to see that the repetitions of attacks on certain formulae were reduced. This fact is in accordance with the principle of the repetition ranks which warrants the finiteness hallmark of the dialogical tableaux.

However, recall that in practice, the rank device enables to perform some constraints such as the time assigned to hold an argumentative debate.

Acknowledgements We are grateful to Prof. Juan Redmond (Universidad de Valparaíso) who motivated our participation in the ISELL (International Symposium of Epistemology, Logic and Language 2012/Lisbon), where the content of this essay was presented for the first time. Many thanks also to the team "Pragmatisme Dialogique", lab. UMR-8163: STL at the University of Lille 3, and particularly to Dr. Nicolas Clerbout (Convenio de Desempeño HACS – Universidad de Valparaíso/Chile) whose **recently published** book on metadialogic has been useful for our own presentation of the dialogical logic. Finally, we deeply thank Prof. Shahid Rahman (Université de Lille, UMR: 8163) for his tireless scientific support and his friendship.

Bibliography

Alchourrón, C.E., Gärdenfors, P., Makinson, D.C.: On the logic of theory change: partial meet contraction and revision functions. J. Symbolic. Logic. **2**(50), 510–530 (1985)

Aristote, O.V.: Topiques. Libraire Philosophique, Paris, J. Vrin, Traduction et notes de J. Tricot, 369 pages (1984)

Aristote: Rhétorique, Paris, éditions Pocket, Coll. « AGORA », Nouvelle traduction du grec, notes et préface de Jean Lauxerois, 287 pages (2007)

Bastin, Y.: « Les langues bantoues », dans Inventaire des études linguistiques sur les pays d'Afrique Noire d'expression française et sur Madagascar, pp. 123–185. In: Barreteau (éd.), CILF, Paris (1978)

Bastin, Y., Coupez, A., et De Halleux, B.: Classification lexicostatistique des langues bantoues (214 relevés), dans Bulletin de l'Académie Royale des Sciences d'Outre-Mer. **XXVII,** 173–199 (1983)

Bench-Capon, T.J.M.: Argument in artificial intelligence and law. Artif. Intell. Law. **5**, 249–261 (1997)

Bondarenko, A., Dung, P.M., Kowalski, R.A., Toni, F.: An abstract, argumentation-theoretic approach to default reasoning. Artif. Intell. **93**, 63–101, Elsevier (1997)

Clerbout, N.: La sémantique dialogique. Notions fondamentales et éléments de métathéorie. College Publications, London, Cahiers de Logique et d'épistémologie"series, vol. 21 (2014)

Clerbout, N., Keiff, L., Rahman, S.: Dialogues and natural deduction. In: Primiero, G., Rahman, S. (ed.) Acts of Knowledge, History, Philosophy, Logic. College Publications, London. chapter 4 (2009)

Dango, A.B.: Approche dialogique de la révision des croyances dans le contexte de la théorie constructive des types. Thése, Lille (2015)

Deschamps, H.: Traditions orales et archives au Gabon. Berger-Levrault, Paris (1962)

Diagne, M.: Critique de la raison orale. Les pratiques discursives en Afrique noire. Karthala, Paris, 600 p (2005)

Dung, P.M.: On the acceptability of arguments and its fundamental role in non-monotonic reasoning, logic programming and N-person games. Artif. Intell. **77**, 321–357, North-Holland Publishing Company (1995)

Felscher, W.: Dialogues as a foundation for intuitionistic logic. In: Gabbay, D., Guenthner, F. (eds.) Handbook of Philosophical Logic, vol. 3, pp. 341–372. Kluwer, Dordrecht (1985)

Fiutek, V., Rückert, H., Rahman, S.: A dialogical semantics for Bonanno's system of belief revision. To appear in Constructions, P. Bour et alii (ed.) College Publications, London (2010)

Fontaine, M., Redmond, J.: Logique Dialogique. Une Introduction. Vol 1 Méthodes de Dialogique : Règles et Exercices. College Publications, Londres, N° 5 de la Série « Cahiers de Logique et d'Épistémologie », 126 p (2008)

Gabbay, D.M., Hogger, C.J., Robinson, J.A. (eds.): Handbook of Logic in Artificial Intelligence and Logic Programming. Volume 3. Nonmonotonic Reasoning and Uncertain Reasoning. Clarendon Press, Oxford (1994)

Gärdenfors, P.: Knowledge in flux. Modeling the Dynamics of Epistemic States. College Publications, Vol. 13 of "Studies in Logic" series, 2008 for the used version (the original one is at 1988), 205 p
Ginsberg, M.L. (ed.): Readings in Nonmonotonic Reasoning. Morgan Kaufmann Publishers, Los Altos (1987)
Hage, J.C.: Reasoning with Rules. An Essay on Legal Reasoning and Its Underlying Logic. Kluwer Academic Publishers, Dordrecht (1997)
Hansson, S.O.: A Textbook of Belief Dynamics. Theory Change and Database Updating. Kluwer Academic Publishers, Dordrecht/Boston/London, 398 p (1999)
Insu S.,Guido G.: A Compact Argumentation System for Agent System specification. STAIRS. In: Peppas, P., Perini,A., Penserini, L. (eds.) IOS Press (2006)
Kamp, H., Reyle, U.: From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory. Kluwer Academic Publishers, Dordrecht/London/Boston (1993), 713 pages
Keiff, L.: Introduction à la dialogique modale et hybride, dans. Philosophia Scientiae. **8**(2), 89–105 (2004)
Keiff, L.: Approches dynamiques à l'argumentation formelle. Ph.D. thesis, Université de Lille, Lille (2007)
Keiff, L.: Dialogical logic, Entry in the Stanford Encyclopaedia of Philosophy. http://plato.stanford.edu/entries/logic-dialogical/ (2009)
Keiff, L., Rahman, S.: La Dialectique entre logique et rhétorique. Revue de Métaphysique et Morale **2**, 149–178 (2010)
Lorenz, K.: Basic objectives of dialogue logic in historical perspective. Synthese **127**, 255–263 (2001)
Lorenzen, P., Lorenz, K.: Dialogische Logik. WBG, Darmstadt (1978)
Rahman, S.: Über Dialogue, Protologische Kategorien und andere Seltenheiten. P. Lang, Frankfurt/Paris/New York (1993)
Rahman, S.: A non normal logic for a wonderful world and more. In: van Benthem, J., et alia (eds.) The Age of Alternative Logics, pp. 311–334, chez. : Kluwer-Springer, Dordrecht (2009)
Rahman, S., Keiff, L.: On how to be a dialogician. In: Vanderveken, D. (ed.) Logic, Thought and Action, pp. 359–408. Kluwer, Dordrecht (2004)
Rahman, S., Rückert, H: New perspectives in dialogical logic. S. Rahman, H. Rückert Synthese. **127**, 1–6 (2001)
Rahman, S., Tulenheimo, T.: From games to dialogues and back: towards a general frame for validity. In: Majer, O., Pietarinen, A.-V., Tulenheimo, T. (eds.) Games: Unifying Logic, Language and Philosophy, Part III. Springer, Dordrecht (2009)
Rahman, S., Clerbout, N., Keiff, L.: Dialogues and natural deduction. In: Primiero, G. (ed.) Acts of Knowledge, History, Philosophy, Logic, pp. 301–336. College Publications, London (2009)
Rückert, H.: Why dialogical logic?. In: Wansing, H. (ed.) Essays on Non-Classical Logic, pp. 165–182. World Scientific Publishing Co. Ltd, Singapore/London (2001)
Suard, F., Buridant, C. (sous la dir.): Richesse du Proverbes. vol 2 : Typologie et Fonctions. Université Lille de III, 275 p (1984)
Tempels, P.: La Philosophie Bantoue. Présence Africaine, Paris (Traduit du Néerlandais par A. Rubbens) 126 pages (1947)
Toulmin, S.: The Uses of Argument. Cambridge University Press, Cambridge, New York (1958)
van Eemeren, F.H., Grootendorst, R., Snoeck, H.F.: Fundamentals of Argumentation Theory. A Handbook of Historical Backgrounds and Contemporary Developments. Lawrence Erlbaum Associates, Mahwah (1996)
Vreeswijuk, G., Prakken, H.: Credulous and Sceptical Argument Games for referred Semantics. In: Proceedings of JELIA'2000, pp. 224–238. Springer (2000)
Wiredu, K. (ed.): A Companion to African Philosophy. Blackwell Companions to philosophy' series, 2004 by Blackwell Publishing Ltd, Malden, MA/Oxford/Victoria, 587 p (2004)

Chapter 15
Semantics of Assertibility and Deniability

Vít Punčochář

Abstract This paper is a reaction to Christopher Gauker's book *Conditionals in Context*. Gauker's semantics of assertibility and deniability will be reconstructed and one peculiar aspect of the semantics will be pointed to: connectives are sensitive to the syntactic structure of the formulas they connect. Even though this is a well motivated principle which can be supported by many examples from natural languages, one of its unwanted consequences is that the resulting formal semantics is not compositional. In this paper, Gauker's semantics will be modified and applied to Stalnaker's concept of context. In this framework, every sentential connective will be replaced with a pair of connectives one of which will be called extensional and the other intensional. This distinction enables us to have an adequate and compositional semantics of assertibility and deniability. We will provide also a syntactic characterization of the logic determined by this semantics.

Keywords Context • Assertibility • Deniability • Pragmatics • Semantics

15.1 Introduction

Christopher Gauker (2005) made an interesting attempt to formulate semantics which is based on the concept of "assertibility in a context" instead of "truth in a world". Gauker's concept of context is defined by recursion: Primitive contexts are consistent sets of literals. Multicontexts of some level are sets which contain primitive contexts and/or multicontexts of lower levels. So, by this recursive definition, we receive an infinite hierarchy of contexts relative to which the relations of "assertibility" and "deniability" are defined.

In this paper, it will be argued that many desirable logical features of Gauker's theory are preserved even if we avoid such complex structures and work simply with Stalnaker's contexts defined as sets of possible worlds (see, e.g., Stalnaker (1999),

V. Punčochář (✉)
Institute of Philosophy and Religious Studies, Faculty of Arts, Charles University in Prague, Prague, Czech Republic
e-mail: vit.puncochar@centrum.cz

J. Redmond et al. (eds.), *Epistemology, Knowledge and the Impact of Interaction*,
Logic, Epistemology, and the Unity of Science 38, DOI 10.1007/978-3-319-26506-3_15

essay 1). And it will be shown that, besides simplicity, this alternative approach has also some other advantages over the original Gauker's approach.

The structure of this paper is as follows. In Sect. 15.2, Gauker's semantics will be presented. We will denote it as S_1. In Sect. 15.3, the semantics S_1 will be simplified in a straightforward way: Gauker's concept of context will be replaced with Stalnaker's simpler concept of context and some consequences of this modification will be discussed. The resulting semantic framework will be denoted as S_2. In Sect. 15.4, it will be shown that a significant drawback of S_1 is the fact that it is not compositional. S_2 does not avoid this problem. In Sect. 15.5, we will further modify S_2 to make it compositional. This final semantics will be denoted as S_3. Some formal features of S_3 will be studied in Sect. 15.6 with the help of one particular modal logic which will be constructed specially for this purpose.

15.2 Gauker's Semantics

The concept of logical consequence is probably the most important concept in logic. According to the traditional Tarskian view, logical consequence is a relation which preserves truth: A conclusion is a logical consequence of premises iff it is impossible for the premises to be true while the conclusion is not true. In this paper, another criterion will be explored according to which logical consequence is a relation which preserves assertibility: A conclusion is a logical consequence of some given premises iff it is impossible for the premises to be assertible while the conclusion is not assertible. Let us call this criterion "Gauker's criterion for the consequence relation". Gauker's book (2005) can be viewed as a thorough investigation of what is the impact of this criterion on logic and the aim of this section is to present his theory.

It is worth mentioning at the beginning that Gauker works with a strict concept of assertibility. If something is assertible in a context the negation of it is strictly excluded by the context. In our semantics this strictness will be preserved. In this respect Gauker's (and our) concept of assertibility differs from the one which was used for example by Ernest Adams (1975), Frank Jackson (1979), Dorothy Edgington (1986) and Jonathan Bennett (2003). For these authors an important aspect of assertibility is high probability but strict necessity is not required.

To provide an informal example illustrating Gauker's concept of assertibility and especially the difference between Gauker's criterion and the standard Tarskian criterion for the consequence relation, we will discuss the famous McGee's counter-example to modus ponens (McGee 1985, p. 462): Consider the situation before the 1980 U.S. presidential election. There were three hot candidates:

Ronald Reagan (a republican).
Jimmy Carter (a democrat).
John Anderson (a republican).

In fact, Reagan won the election, Carter finished second, Anderson third. Consider the following two statements:

A A republican will win the election.
B If it's not Reagan who wins the election, Anderson will win.

Now consider the argument which has the form of modus ponens:

If A, then B. A. Therefore B.

If the full sentences are substituted for *A* and *B*, the argument looks like this:

> If a republican wins the election, then if it's not Reagan who wins it, it will be Anderson. A republican will win the election. Therefore if it's not Reagan who wins, it will be Anderson.

Does this argument together with the real historical scenario provide a natural counterexample to modus ponens? The answer depends on the conception of the consequence relation. If we understand this relation in the standard sense as truth preservation, there is a good reason to regard the argument as a convincing counterexample. If asked, people would probably be strongly inclined to assess the premises as true and the conclusion as false with respect to the described situation. They could provide the following justification: Suppose that *t* is a historical moment just before the election and imagine the scenario of the election. It seems to be reasonable to say that *A* was true at *t* because, in fact, Reagan was later the winner of the election. *If A, then B* was obviously also true at *t* because, in fact, there were just two republican candidates, Reagan and Anderson. However, there is no intuitive sense in which we could regard *B* as true at *t*. On the contrary, *B* seems to be false because the possibility that Reagan would lose the election and Carter (and not Anderson) would win was not excluded at *t*.

A possible objection to this analysis is that it is not clear what the truth conditions for the involved conditionals are. Indeed, serious arguments were formulated supporting the view that ordinary conditionals have no truth conditions at all (see, e.g., Edgington 1986). However, they are objects of our beliefs, something we can be more or less certain about, and something we may or may not be justified to assert. In other words, it is not clear whether conditionals have some semantic value relative to a given state of the world but they certainly have some semantic value relative to a given information state or context. Conditionals maybe do not have truth conditions but they certainly have assertibility conditions.

This is in accordance with McGee. His explanation of the sense in which his "counterexample" invalidates modus ponens differs from the one proposed above: According to McGee, the example shows that "there are occasions on which one has good grounds for believing the premises of an application of modus ponens but yet one is not justified in accepting the conclusion." In McGee's interpretation, entailment is not understood as truth preservation but as rational belief preservation. And his example convincingly shows that modus ponens sometimes does not preserve rational belief.

However, the argument does not fail according to the Gauker's criterion if assertibility is understood in the strict sense. For in every context, in which the

premises are assertible (in the strict sense), the prospect of a win by a democrat has been excluded. In such a context, the conclusion would be assertible as well. (See Gauker 2005, p. 86)

Gauker interestingly noticed that modus tollens version of McGee's example does work as a counterexample even from the (strict) "assertibility" point of view (see Gauker 2005, p. 91): Imagine the same scenario, the same context but different argument:

> If a republican wins, then if Reagan looses the election, Anderson will win. It is not the case that if Reagan looses the election, Anderson will win. Therefore a republican will not win.

This argument is not valid with respect to the Gauker's criterion for logical consequence. In the context of the election, the premises are assertible, but the conclusion is not. Therefore, if we take this example seriously, we should require a logic according to which the inference $p \rightarrow (q \rightarrow r), \neg(q \rightarrow r)/\neg p$ is not classified as logically valid.

While truth is usually regarded as relative to possible worlds, assertibility is relative to contexts. To be able to provide a formal account of logical consequence based on the Gauker's criterion one needs to define assertibility conditions for some formal language, and for this purpose a formal concept of context has to be introduced.

Consider a formal language based on a set of atomic formulas. Literals are atomic formulas and their negations. For Gauker, contexts are certain structures which are built out of literals (so they are linguistic entities). These structures can be defined in the following way. Primitive contexts are nonempty consistent sets of literals. The following sets are defined recursively:

M_0 is the set of all primitive contexts.
$M_{n+1} = M_n \cup (\wp(M_n) - \{\emptyset\})$ where $\wp(M_n)$ is the powerset of the set M_n.
$U = \bigcup_{i=0}^{\infty} M_i$.

Multicontexts are nonempty subsets of U. Contexts are primitive contexts and multicontexts.

Gauker provides us with the following informal idea which is behind the previous definition of primitive contexts:

> The primitive context that actually pertains to the conversation can be approximately defined as the *smallest* formally consistent set of literals such that the interlocutors can reliably be expected to achieve the goal of the conversation if what each of them takes to be the primitive context pertinent to their conversation is that set of literals. (Gauker 2005, p. 13)

Multicontexts are sets of lower level multicontexts or of primitive contexts. The members of multicontexts represent so called "prospects" that a context presents us with. (see Gauker 2005, p. 17)

Gauker works with a formal language which will be denoted as L_1. It contains a set of atomic formulas and complex formulas built out of atomic formulas using the connectives $\neg, \wedge, \vee, \rightarrow$. There is one important restriction in the language L_1: no conditionals occur in antecedents of conditionals.

The letters $p, q, r, \ldots$ will range over atomic formulas. Greek letters $\alpha, \beta, \ldots, \phi, \psi, \chi, \ldots$ will range over all formulas of a language in question. In this section, it is the language L_1. In the following sections also some other languages will be introduced.

A formula is said to be conditional-free if it contains no occurrence of $\rightarrow$. The symbol $\underline{\in}$ stands for "is a member of or is equal to". Now, two relations $\Vdash^+, \Vdash^-$ (assertibility and deniability) between contexts and formulas can be defined. Suppose that C is a context.

Assertibility conditions:

If C is primitive and $p \in C$, then $C \Vdash^+ p$.
If α is a conditional-free formula and $D \Vdash^+ \alpha$ for all $D \in C$, then $C \Vdash^+ \alpha$.
If $C \Vdash^- \phi$, then $C \Vdash^+ \neg\phi$.
If $C \Vdash^+ \phi$ or $C \Vdash^+ \psi$, then $C \Vdash^+ \phi \vee \psi$.
If $C \Vdash^+ \phi$ and $C \Vdash^+ \psi$, then $C \Vdash^+ \phi \wedge \psi$.
If $D \Vdash^+ \psi$ for every $D \underline{\in} C$ such that $D \Vdash^+ \phi$, then $C \Vdash^+ \phi \rightarrow \psi$.
No other formula is assertible in C.

Deniability conditions:

If C is primitive and $\neg p \in C$, then $C \Vdash^- p$.
If α is a conditional-free formula and $D \Vdash^- \alpha$ for all $D \in C$, then $C \Vdash^- \alpha$.
If $C \Vdash^+ \phi$, then $C \Vdash^- \neg\phi$.
If $C \Vdash^- \phi$ and $C \Vdash^- \psi$, then $C \Vdash^- \phi \vee \psi$.
If $C \Vdash^- \phi$ or $C \Vdash^- \psi$, then $C \Vdash^- \phi \wedge \psi$.
If $D \Vdash^- \psi$ for some $D \underline{\in} C$ such that $D \Vdash^+ \phi$, then $C \Vdash^- \phi \rightarrow \psi$.
No other formula is deniable in C.

A formula ϕ is a logical consequence of a set of formulas Δ iff ϕ is assertible in every context in which everything from Δ is assertible. In this case, we also say that Δ / ϕ is a valid form of inference.

Gauker refers to this semantics as "the core theory" and we will denote it as S_1. Gauker himself is not fully satisfied with the logic determined by S_1. Even though it has many features which are in accordance with what Gauker expected from an adequate logic of assertibility and deniability,[1] there are some inference forms which are not valid according to the semantics but which are supposed to be valid according to Gauker's intuitions. Among the most important ones are certainly the following three rules:

1. $p \rightarrow q, \neg q / \neg p$ (modus tollens—the simple case)
2. $\neg p \rightarrow q / p \vee q$ (conditionals-to-disjunctions)
3. $(p \wedge q) \rightarrow r / p \rightarrow (q \rightarrow r)$ (exportation)

Gauker makes some adjustments to make valid these inference patterns. The resulting semantics will not be formulated in this paper. Let us just note that the

[1] E.g., $p \vee q / \neg p \rightarrow q$ and $p, p \rightarrow (q \rightarrow r) / q \rightarrow r$ are valid forms of inference and $p \rightarrow (q \rightarrow r)$, $\neg(q \rightarrow r) / \neg p$ and $p \vee q / (r \rightarrow p) \vee (r \rightarrow q)$ are not—to mention just a few examples.

adjustments seem to be rather artificial and unintuitive and there seems to be no reason for them other than just to gain the validity of some rules including the three above. In the following section, we will introduce a natural modification of Gauker's semantics which will make the three rules of inference valid.

15.3 Stalnaker's Contexts

In this section, Gauker's semantics S_1 will be modified with the help of Stalnaker's concept of context:

> A context should be represented by a body of information that is presumed to be available to the participants in the speech situation. A *context set* is defined as the set of possible situations that are compatible with this information—with what the participants in the conversation take to be the common shared background. (Stalnaker 1999, p. 6)

Gauker explicitly rejects the concept of a possible world. The reason is that this concept presupposes the reference relation and, according to Gauker, so far no one has been able to provide us with a clear account of what the reference relation is (see Gauker 2005, pp. 66–73).

We do not share this skeptical view of possible worlds and hold the view that the concept of a possible world is a technical concept which needs no philosophical justification. Its justification stems purely from its explicative power, i.e., it can be accepted if it enables us to formulate a theoretical framework in which some natural linguistic phenomena can be explicated.

Now the Gauker's original concept of context will be replaced with the Stalnaker's concept of context and some necessary adjustments will be made—e.g., the relation $\Subset$ will be replaced with $\subseteq$.

Let us define possible worlds as classical valuations of atomic formulas, i.e. functions from atomic formulas to the truth values $\{0, 1\}$. Stalnaker's contexts are defined as nonempty sets of possible worlds. We will work with the standard propositional language which contains a set of atoms and the same connectives as L_1, i.e., $\neg, \wedge, \vee, \rightarrow$, which, however, can be arbitrarily nested. This language will be denoted as L_s. The first benefit of our approach is that there seems to be no reason for the restriction which is present in the language L_1. We can form all kinds of nested conditionals and so conditionals can occur also in antecedents of conditionals. In this full language, we are able to formalize also such sentences as *If the light goes on if you press the switch, then the electrician has finished his job*.[2]

We are going to define the relations of assertibility and deniability ($\Vdash^+, \Vdash^-$) between contexts and formulas. It seems to be more comfortable to work with necessary and sufficient conditions instead of only with sufficient conditions (as Gauker does). For this purpose we have to split the set of all formulas into two

[2]The example is taken from (Arló-Costa 1999, p. 8).

classes—conditional-free formulas and formulas which contain an occurrence of the arrow—and for both classes formulate different semantic conditions.

The following semantics is a straightforward modification of Gauker's semantics. Let C be a (Stalnaker's) context. First, suppose that α is conditional-free. Then

$C \Vdash^+ \alpha$ iff for all $w \in C$, α is true in w according to the classical logic.
$C \Vdash^- \alpha$ iff for all $w \in C$, α is false in w according to the classical logic.

Second, suppose that in each of the following conditions, the complex formula is not conditional-free. Then

$C \Vdash^+ \neg\phi$ iff $C \Vdash^- \phi$.
$C \Vdash^- \neg\phi$ iff $C \Vdash^+ \phi$.
$C \Vdash^+ \phi \vee \psi$ iff $C \Vdash^+ \phi$ or $C \Vdash^+ \psi$.
$C \Vdash^- \phi \vee \psi$ iff $C \Vdash^- \phi$ and $C \Vdash^- \psi$.
$C \Vdash^+ \phi \wedge \psi$ iff $C \Vdash^+ \phi$ and $C \Vdash^+ \psi$.
$C \Vdash^- \phi \wedge \psi$ iff $C \Vdash^- \phi$ or $C \Vdash^- \psi$.
$C \Vdash^+ \phi \rightarrow \psi$ iff $D \Vdash^+ \psi$ for all nonempty $D \subseteq C$, such that $D \Vdash^+ \phi$.
$C \Vdash^- \phi \rightarrow \psi$ iff $D \Vdash^- \psi$ for some nonempty $D \subseteq C$, such that $D \Vdash^+ \phi$.

Let us denote this semantics as S_2.[3] The consequence relation can be defined again as preservation of assertibility: Let Δ be a set of formulas and ϕ a formula from L_s. The argument form Δ/ϕ is said to be valid according to S_2 iff its conclusion ϕ is assertible in every context in which all the premises from Δ are assertible. If Δ/ϕ is a valid argument form according to S_2, we will also say that ϕ is a consequence of Δ (in S_2).

It was mentioned as the first advantage of S_2 over S_1 that the language L_s is more expressive than L_1. Second advantage is simplicity. For example, Gauker devotes the whole chapter 9 of Gauker (2005) to the proof of decidability for his core theory. Decidability of S_2 is obtained immediately. If we want to see whether ψ is a consequence of $\phi_1, \ldots, \phi_n$, it is sufficient to check all subcontexts of the context C of all those worlds which assign a truth value only to the atoms occurring in the formulas $\phi_1, \ldots, \phi_n, \psi$. It holds that ψ is a consequence of $\phi_1, \ldots, \phi_n$ iff ψ is assertible in all those subcontexts of C in which $\phi_1, \ldots, \phi_n$ are assertible. Decidability follows from the fact that the number of the subcontexts is finite.

Third advantage is that all important features of Gauker's core theory are preserved and we gained immediate validity of the three forms of inference mentioned above which fail to be valid in the core theory despite Gauker's expectations: the simple case of modus tollens, the inference from conditional to disjunction and the exportation law. All the rules are valid even if we substitute arbitrary conditional-free formulas for atoms.

Let us check the validity of the exportation law to illustrate how the semantics works. The other two rules of inference will play a role in the next section. Suppose that C is a context in which $(p \wedge q) \rightarrow r$ is assertible. We want to prove that in C,

[3] Similar semantic frameworks were discussed also in Punčochář (2013, 2014).

$p \to (q \to r)$ is assertible as well. Suppose that D is an arbitrary subcontext of C such that p is assertible in D. It has to be shown that $q \to r$ is assertible in D. Let E be an arbitrary subcontext of D such that q is assertible in E. Since $E \subseteq D$ and p is assertible in D, p is assertible also in E. So $p \wedge q$ is assertible in E and since E is a subcontext of the context C in which $(p \wedge q) \to r$ is assertible, r has to be assertible in E and that is what we wanted to show.

15.4 Substitution, Replacement of Equivalent Formulas, and Compositionality

In this section, we will point to some peculiarities and unusual aspects of the semantics S_1 and S_2. Almost every formal logical system S studied in the literature has the following two metalogical properties:

(US) The set of valid argument forms in S is *closed under uniform substitution*. That is, if the argument form Δ/ϕ is valid in S and Γ/ψ is the result of substituting an arbitrary formula for every occurrence of an atom in Δ/ϕ, then Γ/ψ is also valid in S.

(REF) The set of valid argument forms in S is *closed under the replacement of equivalent formulas*. That is, if α and β are logically equivalent (i.e. α/β and β/α are both valid argument forms) in S and Δ/ϕ is valid in S and Γ/ψ is the result of replacement of an occurrence of α in Δ/ϕ with β, then Γ/ψ is valid in S.

Neither S_1 nor S_2 has these two properties. The reason is that the semantic systems reflect one surprising phenomenon: In natural languages sentential connectives seem to be sensitive to the syntactic structure of the sentences they connect. This claim is controversial but can be supported by many examples.

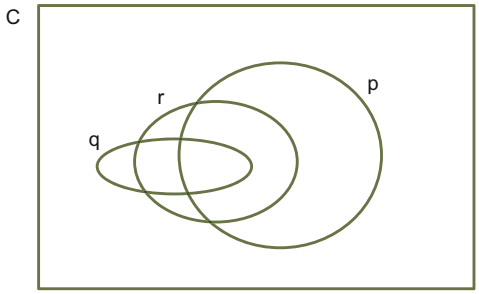
Example 15.1. The first example was discussed already in Sect. 15.2. It was argued there that the scenario of the 1980 U.S. presidential election provided a natural counterexample to the argument:

> If a republican wins, then if Reagan looses the election, Anderson will win. It is not the case that if Reagan looses the election, then Anderson will win. Therefore a republican will not win.

The argument has the modus tollens form where the consequent of the conditional in premises is again a conditional. The fact that the consequent is a conditional sentence is crucial for the counterexample. There is no obvious way how to find such a convincing counterexample to modus tollens in which the consequent would be an elementary non-conditional sentence.

This natural language example is in accordance with our formal semantics S_2. Suppose that C is a context in which $p \to q$ and $\neg q$ are assertible. According to the latter, it holds for every w in C that $w(q) = 0$. Then, according to the former, it has to be the case that for every w in C, $w(p) = 0$. As a result, $\neg p$ is assertible in C.

It has just been proved what was already mentioned in the previous section: The form $p \to q, \neg q / \neg p$ is valid in our semantics. However, when one substitutes $q \to r$ for q one obtains an argument form which is not valid in the semantics. As a consequence, the set of valid arguments is not closed under uniform substitution, i.e. the principle (US) fails in S_2. A counterexample to the rule $p \to (q \to r), \neg(q \to r) / \neg p$ can be illustrated with the following picture:



Notice that in the context C, all q-worlds are r-worlds if we restrict ourselves to p-worlds ($C \Vdash^+ p \to (q \to r)$) but there is a subset of C which contains only q-worlds that are not r-worlds ($C \Vdash^+ \neg(q \to r)$). From this it does not follow that p is false everywhere in C ($C \nVdash^+ \neg p$).

Example 15.2. Consider the following two sentences:

(A) John or David is the murderer.
(B) If John is not the murderer then David is.

They are mutually inferable in the natural language. If one knows that John or David is the murderer, one can definitely infer that if John is not the murderer then David is. And if one knows that it is the case that if John is not the murderer then David is, one can obviously infer that John or David is the murderer. However, the negations of (A) and (B) are not mutually inferable. Consider the situation in which John is among the suspects but David is not. Then one is justified in asserting:

(C) It is not the case that if John is not the murderer then David is.

But one is not justified in asserting:

(D) It is not the case that John or David is the murderer.

It does not follow from the sentence (C) that John is not the murderer but this claim follows from the sentence (D).[4]

Again, this natural language example is in accordance with our formal semantics S_2. One can easily verify that the formulas $p \vee q$ and $\neg p \to q$ are logically equivalent

[4]This example was inspired by R. Stalnaker. See Stalnaker (1999), essay 3, p. 63.

but the formulas $\neg(p \vee q)$ and $\neg(\neg p \rightarrow q)$ are not. However, that also means that the principle (REF) fails in S_2.

Example 15.3. The previous example concerned negation of a conditional. We can illustrate that a similar phenomenon concerns also disjunction of conditionals. Take the following sentences:

(AB) It is the case that Alan is the winner or it is the case that Ben is the winner.
(CD) It is the case that Carl is the winner or it is the case that David is the winner.

As in the Example 15.2, we can notice that on the level of natural language, these two sentences seem to be (respectively) equivalent to the sentences:

(AB)' It is the case that if Alan is not the winner, then Ben is the winner.
(CD)' It is the case that if Carl is not the winner, then David is the winner.

If one knows that (AB), one can infer (AB)'. And if one knows that (AB)', one can infer (AB). Of course, the situation is the same with the sentences (CD) and (CD)'. However, the following two sentences do not seem to be equivalent:

(AD) It is the case that Alan is the winner or it is the case that Ben is the winner or it is the case that Carl is the winner or it is the case that David is the winner.
(AD)' It is the case that if Alan is not the winner then Ben is the winner, or it is the case that if Carl is not the winner then David is the winner.

(AD) says only that one of the four men is the actual winner. (AD)' says that at least one of the conditionals holds. Consider a situation in which there are just four possible winners: Alan, Ben, Carl and David. In such a situation, one could assert (AD), even though none of the disjuncts of (AD) would be assertible. However, (AD)' is not assertible in that situation since none of the two conditionals is assertible. The behavior of disjunction is different in these two cases.

This example is also in accordance with the formal semantics S_2. $p \vee q$, $r \vee s$ are respectively equivalent to $\neg p \rightarrow q$, $\neg r \rightarrow s$ but $(p \vee q) \vee (r \vee s)$ is not logically equivalent to $(\neg p \rightarrow q) \vee (\neg r \rightarrow s)$ so we have here another counterexample to the principle (REF) in S_2.

A lot of examples of this kind can be formulated and they motivate some peculiarities of the semantics S_1 and S_2 such as, e.g., that in these semantics, the behavior of disjunction connecting conditional-free formulas differs from the behavior of disjunction connecting conditionals. In general, the semantic conditions are essentially different for conditional-free formulas on the one side and the rest on the other side. Admittedly, this peculiarity, though motivated by natural language, makes the theory rather inelegant from the technical point of view. The classification of the set of formulas seems to be too rough and too syntactically oriented because even an unimportant occurrence of implication completely changes the semantic status of a formula.

Let us elaborate this point. Two formulas ϕ, ψ are said to be universally interchangeable if in any given formula χ, we can replace any occurrence of ϕ with ψ, and the resulting formula $\chi[\phi/\psi]$ will be always logically equivalent

with χ. The fact that (REF) fails in S_1 and S_2 means that mere logical equivalence (i.e. assertibility in the same contexts) is not a sufficient condition for universal interchangeability. The Examples 15.2 and 15.3 illustrate this fact. The failure of (REF) alone does not lead automatically to the conclusion that the semantics is not compositional, i.e. that the meaning of the whole is not a function of the meanings of its parts and the way they are put together. Logical equivalence in our semantics cannot be understood as the identity of meanings. The reason is that there are two semantically relevant factors in these semantics: assertibility and deniability. For example, the formulas $p \vee q$ and $\neg p \rightarrow q$ are logically equivalent, i.e. assertible in the same contexts, but they are not semantically equivalent in a stronger sense since they are not deniable in the same contexts and hence they differ in meaning. However, if two formulas coincide in both these aspects, that is if they are both assertible and deniable in the same contexts, they should be understood as having the same meaning, i.e. as being semantically equivalent in the stronger sense and, consequently, they should be universally interchangeable. This is the proper sense of the *principle of compositionality* applied to the framework of the semantics S_1 and S_2. Unfortunately, neither S_1 nor S_2 is compositional in this sense which, according to our view, is the main drawback of these semantic systems.

Let us show how the principle of compositionality fails in S_1: For example, the formulas q and $q \wedge (r \rightarrow r)$ are not only assertible but also deniable in the same contexts in S_1. But q and $q \wedge (r \rightarrow r)$ are not universally interchangeable. Let C be the (Gauker's) context $\{\{p\}, \{q\}\}$. Then $C \Vdash^+ p \vee q$ but $C \nVdash^+ p \vee (q \wedge (r \rightarrow r))$. The reason is that there is an arrow occurring in $q \wedge (r \rightarrow r)$ and this mere fact completely changes the semantic status of the formula no matter what the role of the arrow in the formula is. And it seems that in $q \wedge (r \rightarrow r)$ the arrow plays no significant role.

The same situation can be reconstructed also in the semantics S_2. Take the same formulas and (Stalnaker's) context C which contains two worlds v and w such that $v(p) = 1, v(q) = 0, w(p) = 0$ and $w(q) = 1$. The value of r in the two worlds is, say, 1. Then again $C \Vdash^+ p \vee q$ but $C \nVdash^+ p \vee (q \wedge (r \rightarrow r))$.

15.5 Extensional and Intensional Connectives

In this section, we will solve the problem described in the previous section and modify the semantics S_2 to make it compositional. For this purpose, every connective will be replaced by a pair of connectives one of which will be called extensional and the other intensional. Thus instead of one set of connectives, we will work with two sets:

Extensional connectives: $\supset \cap \cup \sim$
Intensional connectives: $\rightarrow \wedge \vee \neg$

Let L_2 be the language containing all atomic formulas and all formulas which can be constructed out of the atomic formulas using the extensional connectives. For the

formulas of the language L_2, the truth and falsity conditions with respect to singular possible worlds are those of classical propositional logic.

Now the language L_3 will be introduced as the smallest set of formulas containing all L_2-formulas and closed under the application of the intensional connectives.

We will use the convention that the Greek letters $\alpha, \beta, \ldots$ will range over the formulas of the language L_2 and the letters $\phi, \psi, \ldots$ over the formulas of the language L_3.

The semantics for this language will be denoted as S_3. Its formulation is almost the same as the formulation of S_2. Only the condition for conditional-free formulas is now replaced with an analogous condition for the formulas from the language L_2. $\Vdash^+$ and $\Vdash^-$ will now stand for the relations of assertibility and deniability between contexts and formulas of the language L_3. The assertibility and deniability conditions are defined in the following way:

$C \Vdash^+ \alpha$ iff for all $w \in C$, α is true in w.
$C \Vdash^- \alpha$ iff for all $w \in C$, α is false in w.
$C \Vdash^+ \neg\phi$ iff $C \Vdash^- \phi$.
$C \Vdash^- \neg\phi$ iff $C \Vdash^+ \phi$.
$C \Vdash^+ \phi \vee \psi$ iff $C \Vdash^+ \phi$ or $C \Vdash^+ \psi$.
$C \Vdash^- \phi \vee \psi$ iff $C \Vdash^- \phi$ and $C \Vdash^- \psi$.
$C \Vdash^+ \phi \wedge \psi$ iff $C \Vdash^+ \phi$ and $C \Vdash^+ \psi$.
$C \Vdash^- \phi \wedge \psi$ iff $C \Vdash^- \phi$ or $C \Vdash^- \psi$.
$C \Vdash^+ \phi \rightarrow \psi$ iff $D \Vdash^+ \psi$ for all nonempty $D \subseteq C$, such that $D \Vdash^+ \phi$.
$C \Vdash^- \phi \rightarrow \psi$ iff $D \Vdash^- \psi$ for some nonempty $D \subseteq C$, such that $D \Vdash^+ \phi$.

Now the consequence relation can be defined again as assertibility preservation.

What is the role of the L_2-formulas? The motivation for their semantics is as follows: A given L_2-formula is assertible (deniable) in a context iff there is enough evidence in the context that the formula is true (false) in the actual world. The actual world is supposed to be one of the possible worlds of the context but from the perspective of the context it is not decided which one it is. So there is enough evidence in the context that the formula is true (false) if and only if the formula is true (false) in all possible worlds of the context. The following result is immediate.

Proposition 15.5.1. *For the formulas of the language L_2 the consequence relation coincides with the consequence relation of classical logic.*

Proof. The proof is trivial. For extensional formulas preservation of assertibility coincides with preservation of truth. □

The relation between extensional and intensional connectives is described in the following three propositions. The first one shows that in the limiting case the semantics of extensional and intensional connectives completely coincide. Suppose that $\phi \in L_3$. ϕ^* will denote the formula from L_2 which is the result of replacing all intensional connectives in ϕ with the corresponding extensional connectives.

Proposition 15.5.2. *For any possible world w, $\{w\} \Vdash^+ \phi$ iff ϕ^* is true in w.*

Proof. Straightforward induction on the complexity of ϕ. □

The next proposition shows that even in a context which contains more than one world the two types of connectives are closely related. We will see that the only exception is disjunction.

Proposition 15.5.3. *Let $\alpha, \beta \in L_2$ and let C be a context. Then*

(a) $C \Vdash^+ \alpha \rightarrow \beta$ *iff* $C \Vdash^+ \alpha \supset \beta$.
(b) $C \Vdash^+ \alpha \wedge \beta$ *iff* $C \Vdash^+ \alpha \cap \beta$.
(c) $C \Vdash^+ \neg\alpha$ *iff* $C \Vdash^+ {\sim}\alpha$.

Proof. For the illustration, we will check the case (a). First, suppose that $C \Vdash^+ \alpha \rightarrow \beta$. We want to prove that there is no world w in the context C such that α is true and β false in w. If there was such a world, we could take the subcontext $D = \{w\}$ of the context C and we would have $D \Vdash^+ \alpha$ but not $D \Vdash^+ \beta$ which is in contradiction with the assumption that $C \Vdash^+ \alpha \rightarrow \beta$.

Second, suppose that $C \Vdash^+ \alpha \supset \beta$. Let D be a subcontext of the context C such that $D \Vdash^+ \alpha$. This means that in every world of D, α is true. It follows that in every world of D, β is true. Therefore $D \Vdash^+ \beta$. We showed that in every subcontext of C, in which α is assertible, β is assertible as well which means that $C \Vdash^+ \alpha \rightarrow \beta$. □

An analogous claim about disjunction does not hold. Therefore, disjunction introduces some divergence between assertibility and truth conditions. However, this connective is not the only source of the divergence, which can be also caused by an appropriate combination of intensional connectives. For example, the formulas $\neg(p \rightarrow q)$ and ${\sim}(p \supset q)$ are not logically equivalent. But if we combine only intensional conjunction with intensional implication the parallel between intensional and extensional connectives still works.

Proposition 15.5.4. *If $\phi \in L_3$ is $\{\neg, \vee\}$-free then $C \Vdash^+ \phi$ iff $C \Vdash^+ \phi^*$.*

Proof. Straightforward induction on the complexity of ϕ. □

Let us now return to natural language and show what is the motivation for the formal semantics S_3 and what kind of phenomena it is supposed to model. The two kinds of connectives—extensional and intensional—in the formal language L_3 do not represent two kinds of connectives in natural language. Instead, they represent two different kinds of usage of the natural language expressions *if*, *or*, *and*, and *not*. When a formula of L_3 is assigned to a sentence of a natural language, it should be taken into account how the logical expressions in the sentence are used, whether extensionally or intensionally.

Let us illustrate the difference on the important case of disjunction. The connective *or* in *A or B* is used extensionally if the sentence says that *A* is true in the actual world or *B* is true in the actual world. And it is used intensionally if the sentence says that *A* is assertible in the given context or *B* is assertible in the given context.

In Adams (1975), Ernest Adams discussed the following natural language argument which is useful for our purposes:

If switches A and B are thrown the motor will start. Therefore, either if switch A is thrown the motor will start or if switch B is thrown the motor will start.

This argument is obviously invalid. Our explanation is that it is invalid because *or* is used intensionally in the conclusion. The conclusion is assertible in a given context (a given space of open possibilities) only if at least one of the disjuncts is.

This is in contrast, e.g., with the sentence "John is in Paris or he is in Prague." This sentence might be assertible even though none of the disjuncts is.

The word *if* is usually used intensionally. The other logical expressions (*or*, *and*, *not*) are usually used extensionally except the cases when they connect sentences in which connectives are used intensionally. Exactly this observation—formulated in different terminology—was directly incorporated in Gauker's semantics S_1 and its modification S_2. However, these semantics do not distinguish two kinds of connectives for two kinds of usage. As we saw in the previous section, the result is that the semantic systems are not compositional. The trick of this section is that the problem is transfered from the formal semantics itself to the process of formalization so that the formal semantics remains technically pure. When intensional and extensional connectives are distinguished, we can add the following instructions how to formalize sentences of natural language:

(a) Formalize *if* as the intensional implication.
(b) As regards *or*, *and*, and *not*, use extensional connectives whenever possible.

For example, consider the sentences from the Example 15.2 of the previous section: *John or David is the murderer* is formalized as $p \cup q$, and *If John is not the murderer then David is* is formalized as $\sim p \rightarrow q$. These two formulas are logically equivalent in S_3, which corresponds to the fact that the two natural language sentences are mutually inferable. Moreover, *It is not the case that John or David is the murderer* is formalized as $\sim(p \cup q)$ but *It is not the case that if John is not the murderer then David is* has to be formalized as $\neg(\sim p \rightarrow q)$ since $\sim(\sim p \rightarrow q)$ is not a well-formed formula of L_3. The formulas $\sim(p \cup q)$ and $\neg(\sim p \rightarrow q)$ are not logically equivalent, which corresponds to the fact that the two formalized sentences are not mutually inferable on the level of natural language as was illustrated in Example 15.2.

In general, if we formalize sentences in accordance with the maxims (a) and (b), and $\phi \in L_3$ is assigned to a natural language sentence then it has to hold that in any formal context C, ϕ is assertible in C according to S_3 iff ϕ^+ is assertible in C according to S_2, where $\phi^+ \in L_s$ is obtained by replacing all the occurrences of extensional connectives with the corresponding intensional connectives. This fact shows in which sense S_3 corresponds to S_2.

The two semantics are intended to model the same phenomena. The difference is purely technical. Unlike S_2, the semantics S_3 is compositional, as will be shown in the rest of this section.

First, notice that (REF) fails also in S_3. E.g., an atom p is equivalent to $(p \vee \neg q) \wedge (p \vee q)$ but $\neg p$ is not equivalent to $\neg((p \vee \neg q) \wedge (p \vee q))$ since the latter formula is equivalent to $(\neg p \wedge q) \vee (\neg p \wedge \neg q)$ which is not assertible in the context $\{v, w\}$ where $v(p) = 0, v(q) = 1, w(p) = 0, w(q) = 0$. Of course, in this context $\neg p$ is assertible.

Nevertheless, unlike S_1 and S_2, the semantics S_3 is compositional. Besides logical equivalence one can define a notion of strong equivalence. We say that two formulas are strongly equivalent (in S_3) if they are not only assertible but also deniable in the same contexts. In other words, two formulas are strongly equivalent if they are logically equivalent and their negations are logically equivalent as well. We will use the symbol $\rightleftharpoons$ for the relation of strong equivalence. The following proposition does not hold for S_2 but it holds for S_3.

Proposition 15.5.5. *Strongly equivalent formulas are universally interchangeable.*

Proof. We can concentrate only on the intensional connectives. The proposition is a consequence of the fact that from the assumption $\phi \rightleftharpoons \psi$ it follows that $\neg\phi \rightleftharpoons \neg\psi$ and also that for an arbitrary formula χ:

$$\phi \wedge \chi \rightleftharpoons \psi \wedge \chi \qquad \chi \wedge \phi \rightleftharpoons \chi \wedge \psi$$
$$\phi \vee \chi \rightleftharpoons \psi \vee \chi \qquad \chi \vee \phi \rightleftharpoons \chi \vee \psi$$
$$\phi \to \chi \rightleftharpoons \psi \to \chi \qquad \chi \to \phi \rightleftharpoons \chi \to \psi$$

□

15.6 Comparison and Syntactic Characterization

Since S_3 is compositional, we prefer this semantics over S_2. In this section, some technical features of the logic determined by S_3 will be explored. The logic will be denoted as $\mathrm{L}(S_3)$. More precisely, $\mathrm{L}(S_3)$ is the set of formulas of L_3 that are assertible in all contexts in the semantics S_3.

It is clear from this paper that the formulation of S_3 was originally motivated by Gauker's ideas. However, it is worth mentioning that it has some common features also with other theories known from the literature: We can mention, e.g., Veltman's data semantics (see Veltman 1986), Nelson's constructive logic (see e.g. Thomason 1969), Wansing's constructive connexive logic (see Wansing 2005) and inquisitive logic (see Ciardelli and Roelofsen 2011). We cannot discuss all these similarities. Instead, we will compare $\mathrm{L}(S_3)$ only with inquisitive logic (InqL) which seems to be, from the technical point of view, most similar to the logic determined by S_3, even though the motivation behind S_3 is different from the motivation behind inquisitive semantics.

Both S_3 and inquisitive semantics provide some kind of intensional semantics for logical connectives. The natural question is, what is the relation of these semantics to modal logic. In this section, a straightforward modal counterpart of inquisitive semantics will be constructed and used as a tool for a syntactic characterization of the logic $\mathrm{L}(S_3)$.

Let us remind some basic notions. The language of modal logic (here denoted as L_m) contains atomic formulas and the formulas built out of atomic formulas using the connectives $\neg, \vee, \wedge, \to, \Box$. $\Diamond$ is an abbreviation for $\neg\Box\neg$. So L_m is an extension of the language L_s by the modal operator $\Box$. A Kripke model is a triple

$M = \langle W, R, V\rangle$ where W is an arbitrary nonempty set, R is a binary relation on W and V is a function assigning a subset of W to every atomic formula.

The relation of truth ($\vDash$) between the elements of W of a Kripke model $M = \langle W, R, V\rangle$ and formulas from L_m is defined in the following way:

$M, w \vDash p$ iff $w \in V(p)$, for every atomic formula p.
$M, w \vDash \phi \wedge \psi$ iff $M, w \vDash \phi$ and $M, w \vDash \psi$.
$M, w \vDash \phi \vee \psi$ iff $M, w \vDash \phi$ or $M, w \vDash \psi$.
$M, w \vDash \phi \rightarrow \psi$ iff $M, w \nvDash \phi$ or $M, w \vDash \psi$.
$M, w \vDash \neg\phi$ iff $M, w \nvDash \phi$.
$M, w \vDash \Box\phi$ iff $M, v \vDash \phi$ for all v such that wRv.

An intuitionistic Kripke model is a Kripke model $M = \langle W, R, V\rangle$ where R is a partial order (that is a reflexive, antisymmetric, and transitive relation) on W and V is persistent: if $w \in V(p)$ and wRv then $v \in V(p)$. The intuitionistic relation of truth ($\vDash_i$) between the elements of W of an intuitionistic Kripke model $M = \langle W, R, V\rangle$ and formulas from L_s is defined in the following way:

$M, w \vDash_i p$ iff $w \in V(p)$, for every atomic formula p.
$M, w \vDash_i \phi \wedge \psi$ iff $M, w \vDash_i \phi$ and $M, w \vDash_i \psi$.
$M, w \vDash_i \phi \vee \psi$ iff $M, w \vDash_i \phi$ or $M, w \vDash_i \psi$.
$M, w \vDash_i \phi \rightarrow \psi$ iff $M, w \vDash_i \psi$ for all v such that wRv and $M, v \vDash_i \phi$.
$M, w \vDash_i \neg\phi$ iff $M, v \nvDash_i \phi$ for all v such that wRv.

Let us define the m-logic of a Kripke model $M = \langle W, R, V\rangle$ as the set of formulas ϕ from L_m such that $M, w \vDash \phi$ for all $w \in W$. The i-logic of an intuitionistic Kripke model $M = \langle W, R, V\rangle$ is defined as the set of formulas ϕ from L_s such that $M, w \vDash_i \phi$ for all $w \in W$.

Let B be a set of atomic formulas. Then $M^B = \langle \wp(W^B) - \{\emptyset\}, \supseteq, V^B\rangle$ denotes the intuitionistic Kripke model where W^B is the set of all possible worlds with respect to B, i.e. the set of all functions from B to the truth values $\{0, 1\}$, $\supseteq$ is the superset relation, and V^B is defined as follows: if $p \in B$ then $V^B(p) = \{C \in \wp(W^B) - \{\emptyset\}; \text{for all } w \in C, w(p) = 1\}$ and if $p \notin B$ then $V^B(p) = \emptyset$.

As in Ciardelli and Roelofsen (2011), inquisitive logic InqL can be defined as the i-logic of the model M^A where A is the set of all atomic formulas of L_s.

Now, we will explore the m-logic of the model M^A. Let us denote the logic as mInqL. A simple semantic observation shows that mInqL is a modal companion of InqL which means that these logics are related via the well-known Gödel's translation g from the language L_s to the language L_m defined as follows:

$g(p) = \Box p$ for every atomic formula p.
$g(\neg\phi) = \Box\neg g(\phi)$.
$g(\phi \wedge \psi) = g(\phi) \wedge g(\psi)$.
$g(\phi \vee \psi) = g(\phi) \vee g(\psi)$.
$g(\phi \rightarrow \psi) = \Box(g(\phi) \rightarrow g(\psi))$.

The relation between InqL and mInqL is spelled out in the following theorem.

Theorem 15.6.1. *$\phi \in$ InqL iff $g(\phi) \in$ mInqL, for every $\phi \in L_s$.*

It is evident which translations can be used if one wants to relate in an analogous way the logic InqL to L(S_3) and L(S_3) to mInqL.

To articulate the relation of InqL to L(S_3) we will introduce a translation u from L_s to L_3:

$u(p) = p$ for every atomic formula p.
$u(\neg\phi) = u(\phi) \rightarrow (p \wedge \neg p)$.
$u(\phi \wedge \psi) = u(\phi) \wedge u(\psi)$.
$u(\phi \vee \psi) = u(\phi) \vee u(\psi)$.
$u(\phi \rightarrow \psi) = u(\phi) \rightarrow u(\psi)$.

Theorem 15.6.2. *$\phi \in$ InqL iff $u(\phi) \in L(S_3)$, for every $\phi \in L_s$.*

Another translation t, now from L_3 to L_m is defined in the following way:

$t(\alpha) = \Box\Diamond\alpha^+$ for every $\alpha \in L_2$.[5]
$t(\neg\alpha) = \Box\Diamond\neg\alpha^+$ for every $\alpha \in L_2$.
$t(\neg\neg\phi) = t(\phi)$.
$t(\phi \wedge \psi) = t(\phi) \wedge t(\psi)$.
$t(\neg(\phi \wedge \psi)) = t(\neg\phi) \vee t(\neg\psi)$.
$t(\phi \vee \psi) = t(\phi) \vee t(\psi)$.
$t(\neg(\phi \vee \psi)) = t(\neg\phi) \wedge t(\neg\psi)$.
$t(\phi \rightarrow \psi) = \Box(t(\phi) \rightarrow t(\psi))$.
$t(\neg(\phi \rightarrow \psi)) = \Diamond(t(\phi) \wedge t(\neg\psi))$.

The following theorem relates our logic L(S_3) to mInqL.

Theorem 15.6.3. *$\phi \in L(S_3)$ iff $t(\phi) \in$ mInqL, for every $\phi \in L_3$.*

In the rest of this section, we will axiomatize the logic mInqL and by that we will obtain an indirect axiomatization of L(S_3). A logic similar to mInqL was introduced in Punčochář (2012) under the name L(E1). The logic L(E1) was interpreted in a natural way as an epistemic modification of Carnap's modal logic *C*. This interpretation does not work for mInqL which, however, seems to be more appropriate for our present purposes. mInqL will serve us only as a technical tool and we do not intend to provide a natural interpretation of this logic. The structure of the completeness proof for mInqL is the same as the one used in Punčochář (2012) for L(E1). A similar completeness proof for inquisitive logic was also formulated in Ciardelli and Roelofsen (2011).

Let ϕ, ψ and $\phi_1, \ldots, \phi_n$ range over the formulas from L_m. p will range over atomic formulas. Consider the following Hilbert calculus which contains (mp) and (nec) as its inference rules plus the axiomatic schemas (ClT), (K), (4), (X1), (X2) and (Yn) for every natural number n:

[5]Let us remind that α^+ is obtained from α by replacing all the occurrences of extensional connectives with the corresponding intensional connectives.

All classical tautologies	(ClT)
$\Box(\phi \to \psi) \to (\Box\phi \to \Box\psi)$	(K)
$\Box\phi \to \Box\Box\phi$	(4)
$p \to \Box p$	(X1)
$\Box\Diamond p \to p$	(X2)
$\bigwedge_{i=1}^{n} \Diamond\Box\phi_i \to \Diamond(\bigwedge_{i=1}^{n} \Diamond\Box\phi_i \wedge \Box\Diamond \bigvee_{i=1}^{n} \phi_i)$	(Yn)

Let us denote this calculus as C. We will prove that C is sound and complete with respect to mInqL. For this purpose the following proposition will be useful.

Lemma 15.6.1. *If A is the set of all atomic formulas and $B \subseteq A$, then it holds for every formula $\phi \in L_s$ which contains only atoms from B that ϕ is in the i-logic of M^B iff ϕ is in the i-logic of M^A. Analogously, if $\phi \in L_m$ and ϕ is built out of atoms from B then ϕ is in the m-logic of M^B iff ϕ is in the m-logic of M^A.*

So if we want to know if ϕ is in InqL or in mInqL, it suffices to decide whether ϕ is in the i-logic or m-logic of the finite model M^B where B is the set of atomic formulas from ϕ. As a consequence, InqL and mInqL are decidable.

We will also need the concept of p-morphism. Let $M_1 = \langle W_1, R_1, V_1\rangle$, $M_2 = \langle W_2, R_2, V_2\rangle$ be two Kripke models and B a set of atomic formulas. A function f from W_1 to W_2 is called a p-morphism from M_1 to M_2 with respect to B iff the following three conditions are satisfied.

1. for every $w \in W_1$ and $p \in B$, $w \in V_1(p)$ iff $f(w) \in V_2(p)$.
2. for every $w, v \in W_1$, if wR_1v then $f(w)R_2f(v)$.
3. for every $w \in W_1$ and $t \in W_2$, if $f(w)R_2t$ then there is $v \in W_1$ such that wR_1v and $f(v) = t$.

The following lemma states a well-known fact that p-morphism is a truth-preserving function (see, e.g., Chagrov and Zakharyaschev 1997).

Lemma 15.6.2. *Let $M_1 = \langle W_1, R_1, V_1\rangle$, $M_2 = \langle W_2, R_2, V_2\rangle$ be two Kripke models. Suppose that f is a p-morphism from M_1 to M_2 with respect to a set of atomic formulas B. Suppose that $w \in W_1$ and $\phi \in L_m$ is built out of atoms from B. Then $M_1, w \vDash \phi$ iff $M_2, f(w) \vDash \phi$. If M_1 and M_2 are intuitionistic Kripke models and $\phi \in L_s$ then also $M_1, w \vDash_i \phi$ iff $M_2, f(w) \vDash_i \phi$.*

A Kripke model M is a model of C if every formula provable in C is contained in the m-logic of M. We say that a schema is valid in M if every instance of the schema is contained in the m-logic of M. The following lemma is crucial for the completeness proof.

Lemma 15.6.3. *If M is a Kripke model of C and B is a finite set of atoms then there is a p-morphism from M to M^B with respect to B.*

Proof. Let $M = \langle W, R, V\rangle$ be a Kripke model of C and $B = \{p_1, \ldots, p_m\}$ a finite set of atoms. For any $w \in W$ let $\bar{w}$ denote the function from B to the truth values such that for every $p \in B$, $\bar{w}(p) = 1$ iff $w \in V(p)$. We say that $v \in W$ is final iff for every w, if vRw then $\bar{w} = \bar{v}$.

The function f from W to W^B is defined as follows:

$f(w) = \{\bar{v}; wRv \text{ and } v \text{ is final}\}$.

Notice that (X1) guarantees that $f(w)$ must be nonempty. We have to show that f is a p-morphism, so we have to check that the conditions 1-3 hold.

1. Suppose that $p \in B$. Since (X1) is valid in M, it holds that if $M, w \vDash p$ then $M^B, f(w) \vDash p$. Since (X2) is valid in M, we have that if $M^B, f(w) \vDash p$ then $M, w \vDash p$.
2. Suppose that wRv. Since (4) is valid in M, we have $f(v) \subseteq f(w)$ as required.
3. Suppose that $t \subseteq f(w)$. Then there are some final $v_1, \ldots v_n \in W$ such that $t = \{\bar{v_1}, \ldots \bar{v_n}\}$ and wRv_1 and $\ldots$ and wRv_n. For every $v \in W$, let χ_v be the formula $l_1 \wedge \ldots \wedge l_m$ where l_i $(1 \leq i \leq m)$ is p_i if $v \in V(p_i)$ and $l_i = \neg p_i$ if $v \notin V(p_i)$. Then $M, w \vDash \bigwedge_{i=1}^{n} \Diamond\Box\chi_{v_i}$. Since (Y$n$) is valid in M, $M, w \vDash \Diamond(\bigwedge_{i=1}^{n} \Diamond\Box\chi_{v_i} \wedge \Box\Diamond \bigvee_{i=1}^{n} \chi_{v_i})$. Therefore, there is $v \in W$ such that wRv and $M, v \vDash \bigwedge_{i=1}^{n} \Diamond\Box\chi_{v_i} \wedge \Box\Diamond \bigvee_{i=1}^{n} \chi_{v_i}$. It follows that $f(v) = t$.

□

Now we are prepared to prove the completeness theorem.

Theorem 15.6.4. *A formula from L_m is provable in C iff it is in mInqL.*

Proof. The system is sound with respect to mInqL as can be easily verified. For the completeness part, suppose that ϕ is not provable in C. Since C is an extension of the logic K (contains (ClT), (K), (mp), (nec)), we can use a basic result from modal logic and conclude that there is a Kripke model $M = \langle W, R, V\rangle$ of C and $w \in W$ such that $M, w \nvDash \phi$. According to Lemma 15.6.3, there is a p-morphism from M to M^B with respect to B where B is the set of atoms occurring in ϕ. Then, according to Lemma 15.6.2, $M^B, f(w) \nvDash \phi$. So ϕ is not in the m-logic of M^B. Therefore, according to Lemma 15.6.1, ϕ is not in the m-logic of M^A, i.e. $\phi \notin$ mInqL. □

Due to Theorem 15.6.3, the axiomatization of mInqL provides also a syntactic characterization of our logic L(S_3).

15.7 Conclusion

In this paper we introduced the Gauker's semantics according to which the consequence relation is a relation which preserves assertibility rather than truth. We applied directly Gauker's approach to the Stalnaker's concept of context. Then we identified a problem connected with both Gauker's original theory and our straightforward modification of the theory. The problem was that these theories violate the principle of compositionality. We proposed another modification S_3 which is based on the distinction between extensional and intensional connectives. The semantics S_3 is compositional and so avoids the problem of the former theories.

In S_3, it is possible to capture all the important phenomena which the previous semantics deal with. For example, if we want to reflect the fact that disjunctions of conditionals often behave in a different way than disjunctions of elementary sentences (the attempt to capture such phenomena caused the failure of compositionality in Gauker's theory), we can simply formulate some maxims which regulate the process of formalization. The disjunctions of elementary sentences can be systematically formalized as extensional disjunctions and disjunctions of conditional sentences can be formalized as intensional disjunctions. Therefore nothing is lost and semantic purity is gained.

At the end, we studied S_3 from the technical point of view and provided a syntactical characterization of the logic determined by this semantics.

Acknowledgements Work on this paper has been supported by grant no. P401/11/0371 of the Czech Science Foundation.

References

Adams, E.W.: The Logic of Conditionals. An Application of Probability to Deductive Logic. D. Reidel, Dordrecht (1975)

Arló-Costa, H.: Belief revision conditionals: *Basic* iterated systems. Ann. Pure Appl. Log. **96**, 3–28 (1999)

Bennett, J.: A Philosophical Guide to Conditionals. Oxford University Press, Oxford (2003)

Chagrov, A., Zakharyaschev, M.: Modal Logic. Oxford University Press, Oxford (1997)

Ciardelli I., Roelofsen, F.: Inquisitive logic. J. Philos. Log. **40**, 55–94 (2011)

Edgington, D.: Do conditionals have truth-conditions? Critica **18**, 3–30 (1986)

Gauker, Ch.: Conditionals in Context. MIT Press, London (2005)

Jackson, F.: On assertion and indicative conditionals. Philos. Rev. **88**, 565–589 (1979)

McGee, V.: A counterexample to Modus Ponens. J. Philos. **82**, 462–471 (1985)

Punčochář, V.: Some modifications of Carnap's modal logic. Stud. Log. **100**, 517–543 (2012)

Punčochář, V.: Pravdivost vs. tvrditelnost. Organon F, 20, Supplementary Issue 1, 122–143 (2013)

Punčochář, V.: Intensionalisation of logical operators. In: Dančák, M., Punčochář, V. (eds.) The Logica Yearbook 2013, pp. 173–186. College Publications, London (2014)

Stalnaker, R.C.: Context and Content. Oxford University Press, Oxford (1999)

Thomason, R.: A semantical study of constructive falsity. Zeitschrift für mathematische Logik und Grundlagen der Mathematik **15**, 247–257 (1969)

Veltman, F.: Data semantics and the pragmatics of indicative condicionals. In: Traugott, E.C., Meulen, A., Reilly, J.S., Ferguson, Ch.A. (eds.) On Conditionals, pp. 147–169. Cambridge University Press, Cambridge (1986)

Wansing, H.: Connexive modal logic. In: Schmidt, R. et al. (eds.) Advances in Modal Logic, vol. 5, pp. 367–383. King's College Publications, London (2005)

Chapter 16
The Quest for the Concept in the Twentieth Century: Predicates, Functions, Categories and Argument Structure

Francisco J. Salguero-Lamillar

Abstract Philosophers, logicians, linguists and even mathematicians have tried to decipher the mechanisms by which concepts are constructed from the meaning of words. One way to achieve this was the study of lexical meaning and its combinatorial properties. Our purpose is to explore the seminal ideas that have resulted in categorial grammars and their relationship with other grammatical models and actual theories of meaning, in a historical process that takes us from the notion of category to that of predication, and from this to the notion of function, then to functional categories and finally to the linguistic notion of argument structure.

Keywords Natural language • Conceptualization • Categories. Categorial grammar • Unification • Argument structure

16.1 Introduction

Natural Language is the most powerful tool for generating complex concepts. Actually, it is not the only tool for the conceptual tasks of the mind, but we can hypothesize that human brain could not reach the same level of complexity without the support of natural language. Perception, memory, cultural judgements are behind the constitution of any complex representation in the mind, but we can prove that these mental representations need to be grouped and categorized by a linguistic label that relates them to other representations as another element of a (complex) system. This explains what we can denominate *learning transfer*, based on the psychological notion *transfer of practice,* first introduced at early twentieth century by Thorndike and Woodworth (1901). The process of learning involves achieving mental representations capable of relating or influencing other representations by similarity or analogy. To achieve this, words are essential.

F.J. Salguero-Lamillar (✉)
Group of Logic, Language and Information, University of Seville, Seville, Spain
e-mail: salguero@us.es

J. Redmond et al. (eds.), *Epistemology, Knowledge and the Impact of Interaction*,
Logic, Epistemology, and the Unity of Science 38, DOI 10.1007/978-3-319-26506-3_16

So, when by the end of the twentieth century Kanerva (1988) tried to modelize computational memory by means of his idea of a *sparse distributed memory*, he realized that human memory has a tendency to congregate memories related by similarities due to personal experience, culture and language skills. To retrieve a concept or a piece of memory, we must implement a series of mental skills supported by *perceptual attractors*, the most important being the standard lexical items, arranged as a mental lexicon.

In a similar way, when American anthropologists wanted to explain the differences in culture and worldview—*Weltanschauung*—among American Indians and the European culture at the beginning of the twentieth century, they also realized that the linguistic structures interfere at all levels with the way reality is perceived by the speakers of a certain language. This was the origin of the relativistic hypothesis called *Sapir-Whorf Hypothesis*, which can be found in its most known formulation by Benjamin Whorf:

> We are thus introduced to a new principle of relativity, which holds that all observers are not led by the same physical evidence to the same picture of the universe, unless their linguistic backgrounds are similar, or can in some way be calibrated. (Whorf 1956: 214)

In short, what this means is that mental functions, memory retrieval, and the perception of the world, as well as the interpretation of reality, have been linked directly to our language skills throughout the twentieth century by psycologists and linguists—besides the well-known and widely described elsewhere *linguistic turn* in analytic philosophy—, postulating the dependence of mental representations and conceptualization capabilities on our language faculty.

However, though the relation among concepts and words could look very natural, it is not a biunivocal relationship at all. In every human language we can have many-to-one as well as one-to-many relations among words and concepts. Of course, we are referring to the cases of synonymy and polysemy. Both phenomena are very interesting from the point of view of the construction of mental lexicon, because they point at grammar and not only at lexical semantics.

So, it is impressive the fact that in all languages it is possible to express the same concepts or ideas in several grammatical ways. The existence of synonymous expressions is a universal phenomenon that seems to contradict the supposed principle of "linguistic economy". However, the fact that we can say the same using different words or linguistic structures—such as phrases and phraseological units—makes natural language maximally expressive and creative, and allows a greater amount of expressive resources for communicative purposes, building mental bridges among different semantic and cognitive fields.

On the other hand, the existence of polysemy is also a linguistic universal. At first glance, polysemy may be seen as an excessive resource of natural languages that leads to ambiguity and, therefore, to malfunction. On the contrary, polysemous words act as super-connectors in a complex network formed by different semantic subnets (the wordnet), and give cohesion to the complete network, making navigation and local association of the underlying concepts easier. In other words, "polysemy organizes the semantic graph in a compact and categorical representation, in a way that may explain the ubiquity of polysemy across languages" (Sigman and Cecchi 2002).

As a complement to the mental lexicon, it is also possible in all languages expressing concepts by means of structural rules that use more than one single word. This is the case of the concepts that are expressed by phraseological units—more or less opaque semantically—as "kick the bucket" (DIE) or "son of my brother" (NEPHEW), for instance. They may not necessarily be always identified by a single lexical item, but, in any case, they can be part of the argument structure of some sentence: "the son of my brother's best friend studies astrophysics in Canada". The combinatorial rules that allow build new concepts from simpler concepts by means of phrase structure and argument structure are those that give the language its unlimited ability to represent reality, beyond perception and memory, and beyond the mere symbolism of individual isolated words.

So, achieving a formal logical representation model of linguistic mechanisms to build complex concepts from lexical meaning has been in the past and is at present one of the great challenges of linguistics, logic and computation theory, especially in the last hundred years.

16.2 Categories and the Science of Meanings

16.2.1 Antecedents

We can discuss whether the first antecedent of the compositional theories of meaning is found in Plato (*Sophist* 261d–263b), when he says that only after the union of a name (*onoma*) and a verb (*rhema*) discourse is given and it is possible to ascribe truth or falsity to the corresponding sentences. Nevertheless, we must agree it was the Aristotelian distinction among the parts of speech and their combinatorial properties to get predication the most influential, through syllogistic logic, in the subsequent theories about meaning and compositionality.

The categorial classification of simple expressions of language made by Aristotle in his treatise on *Categories* is based on semantic and conceptual criteria, rather than morphological or functional ones, as will be the case later with the stoic clasification of the different parts of speech—a hybrid between semantic and morphosyntactic criteria as case inflection in nouns or conjugation in verbs—and the modifications of the stoic categories made later by the Alexandrian grammarians Dionysius Thrax (first century BC) and Apollonius Dyscolus (second century AD).[1]

[1]Functional and morphosyntactic categories that reach the Roman grammarians—and from them the early modern period—are those proposed by the Alexandrian, who in turn adapted the categories proposed by the Stoic school. The Stoics used a semantic criterion for the classification of the parts of speech, following the Aristotelian style, but they introduced functional changes based on the morphology of words. So, they made a distinction in Aristotle's *syndesmoi* between a group of words with inflection that they called *arthra* (pronouns and articles), and the invariable words (prepositions and conjunctions); they also created the category of adverbs (*mesotes*) based on syntactic criteria—adverbs appear combined with the verb—as well as morphological ones—

Aristotle's *Categories* studies the range of variability of subject and predicate. The treatise begins distinguishing sign and meaning through the study of univocal (synonymy), equivocal (homonymy and polysemy) and paronymical words (derivative meanings). But the key to the Aristotelian theory is, no doubt, the idea of a composition of meanings based on predication as the main function for getting complex expressions by combining categories. Predication results on sentence structure, which is the basis of the enunciative functions of language, where the concepts of truth and falsity are built on. So, the categories will be later conceived in the Middle Ages as *praedicamenta* (predicates); and a predicate is, following Aristotle, what is said *of* a subject or even what is *in* a subject, i.e.: an inherent concept to a given meaning—in the different ways in which predicates will be classified afterwards in the treatise on *Topics*.

> Of things that are said, some involve combination while others are said without combination. Examples of those involving combination are: man runs, man wins; and of those without combination: man, ox, runs, wins. [...] Whenever one thing is predicated [*kategoretai*] of another as of a subject, all things said of what is predicated will be said of the subject also. For example, man is predicated of the individual man, and animal of man; so animal will be predicated of the individual man also—for the individual man is both a man and an animal. [...] Of things said without any combination, each signifies either substance or quantity or qualification or a relative or where or when or being-in-a-position or having or doing or being-affected. [...] None of the above is said just by itself in any affirmation, but by the combination of these with one another an affirmation is produced. For every affirmation, it seems, is either true or false; but of things said without any combination none is either true or false (e.g. man, white, runs, wins). (*Categories*, 1a16–2a5, translated by J. L. Ackrill)

These ideas germinated in the Middle Ages in the so-called *Speculative Grammars* or treatises *De modis significandi*, and the concept of a Universal Grammar based on Aristotelian logic. The foundation of these philosophical theories about language is the reinterpretation of predicables in three different ways, as *modi essendi* (modes of existence), *modi intelligendi* (modes of conceptualization) and *modi significandi* (modes of signification). This distinction resulted in the designation of these philosophers who were known as *Modistae*.[2] The *modi essendi* were widely intepreted as *modus entis* (when they refer to permanent properties of beings) and *modus esse* (when they refer to temporal changes and mutation processes of things). Mind apprehends the knowledge of existing things through *modi intelligendi*

adverbs are formed on nominal or adjectival themes and roots. Finally, Stoics introduced the concept of *klisis* to denote the grammatical variation of a word, limiting the aristotelian notion of case (*ptosis*) to the words of nominal category (proper names, common nouns and adjectives) as well as to the words classified as *arthra* (pronouns and articles), this being the basis for the distinction between these categories and verbs (*rhema*). A more extensive description of the topic can be found at Robins (1969).

[2] Among them we can highlight Martin of Dacia, Michel de Marbais, Peter Helias and Thomas of Erfurt. The philosophy of the *Modistae* drank directly from the work of Aristotle, but also had the probable influence of commentators such as Duns Scotus. Their theories were also clearly related to those of other philosophers of the period as Roger Bacon or William of Ockham.

activi and the knowledge of its transient properties using the *modi intelligendi passivi*. Subsequently, this knowledge of existing things is expressed by the *modi significandi activi*, and the knowledge of their qualities by the *modi significandi passivi*.

The notion of *modi significandi* is key to understanding the philosophical system of *Modistae*. All the parts of speech represent reality in a certain way, so that the categories are interpreted as *modi* with its own semantic component. Furthermore, it is the semantic component what allows the words to appear in predication, either as *suppositum* or as *appositum*, i.e.: as subject or predicate.

One of the most important disputes within this tradition was that of the semantics of universal terms. Universals are general terms (sometimes abstract terms) as "man", "truth", "beauty", "being", and so on. The controversy revolved around the fact that such terms could be both subjects and predicates in the statements that constitute the Aristotelian syllogism, so it was not clear to which category they belonged, and, in the case they were conceived as substances, if their way of being and their way of meaning were the same as those of terms such as "Socrates", "something true", "something beautiful" or "something that is or exists". The *Modistae* held a position with respect to this question that has been called *moderate realism*. They thought that universals are abstracted from the properties of real things, changing from *modi essendi* to *modi significandi* through *modi intelligendi*—unlike the nominalists, for whom universals had only a mode of signification (i.e.: they were considered just words that refer to all human beings and all that is true or beautiful or existing, respectively).

These theories about universal grammar and the arising of general concepts from the cognitive interpretation of the categories, starting from their ontological and logico-linguistic interpretation, influenced the appearance during the modern period of *rationalist grammars*—J. C. Scaligero's *De causis linguae latinae libri XIII* (1540), Francisco Sánchez de las Brozas' *Minerva sive de causis linguae latinae* (1587) or *Port-Royal Grammar* (1660)—and serves as an explanation of the origin of the Leibnizian project for a *mathesis universalis*—G. W. Leibniz: *De arte combinatoria* (1666).

The main features of the *mathesis universalis* project are its goal—a universal science consisting on symbols—, its method—based on the linguistic analysis from complex terms to their most simple "formal parts" (indefinable terms)—, and the tools to achieve it—mathematical symbols that represent these formal parts and a few rules for their combination that must be given.

These indefinable mathematical symbols and the rules for combining them would describe a universal logic of meaning and discovery, that is: a procedure for reaching new concepts and meanings from those facts and truths already known.[3] The door to a general mathematical science of significations was open.

[3]Leibniz's proposal was very important and influential on the search for universal languages during the eighteenth Century, as well as were his etymological studies on the rise of comparative linguistics at the nineteenth Century.

16.2.2 The Phenomenological Origin of Categorial Grammar: Husserl's General Science of Significations

In this historical and theoretical context is where we can understand Husserl's theories about a *general science of significations*, as was conceived in the early twentieth century:

> The task of an accomplished science of meanings would be to investigate the law-governed, essence-bound structure of meanings and the laws of combination and modification of meaning which depend upon these, also to reduce such laws to the least number of independent elementary laws. We should obviously also need to track down the primitive meaning-patterns and their inner structures, and, in connection with these, to fix the pure categories of meaning which circumscribe the sense and range of the indeterminates—the 'variables' in a sense exactly analogous to that of mathematics—that occur in such laws. (Husserl (1900–1901): *Fourth Logical Investigation*, §13)

Edmund Husserl proposes here a concept of grammar based on a priori laws that determine language meaning rather than "exclusively on psychology and other empirical disciplines". That is to say, he returns to "the old idea of a general grammar" instead of the new empiricist trends in linguistics of the newly opened century. This is a semantic perspective, in the sense that for him linguistic expressions are significations (*Bedeutungen*) that are assigned semantic categories (*Bedeutungskategorien*). These significations can be simple (the lexicon) and compound (the sentences), so that simple significations are only partial meanings that require completion to give full meanings using certain combinatorial rules.

In this respect, when Husserl wonders whether syncategorematic elements are significant elements of complex expressions, he admits that their meaning is not the same as that of the categorematic elements, although both kinds of significations really become meaningful only when they are complemented to form a compound expression. In other words, their grammatical distinction admits another interpretation: conceiving the integrity, or partiality, of the expressions as resulting from certain integrity, or partiality, of the significations, i.e.: "the grammatical distinction as a result of some essential difference of meaning" (*Fourth Logical Investigation*, §4). Therefore, Husserl does not admit that syncategorematic linguistic elements lack of meaning, but he holds that the difference with respect to the categorematic elements is "some essential difference of meaning" to be determined. From this regard emerges the distinction between independent and non-independent significations, related to the previous distinction between dependent and independent objects made in the *Third Logical Investigation* "On the theory of wholes and parts". These distinctions are phenomenological, based on the concept of understanding, and go beyond grammar or logic.

We can compare Husserl's proposal with Frege's in order to understand the similarities and also the specific features that differentiate both conceptions of complex meaning. The well-known Frege's Principles of Compositionality and Context can be stated as follows:

- *Principle of Compositionality*: The meaning of a complex expression is a function on the meaning of the most simple expressions that compound it and the rules of combination used over these simple expressions to generate the complex one.[4]
- *Context Principle*: It is necessary to consider the words as a part of the sentence when we ask for their meaning, and it is enough when a whole complete sentence has a meaning, because thereby also its parts receive their content.[5]

Unlike Frege's view, Husserl's proposal is not logicist; i.e.: there is no identification between the laws of logic and the rules of grammar. Nevertheless, for him the rules of grammar are logical, what means that there is a (certain kind of) logic behind the combinatorial rules of significations. Every language works on the basis of a general logic that establishes laws of possibility and exclusion, and these laws—and not others—make up what he calls *pure grammar*. Hence, pure logical grammar is conceived as a set of analytical laws common to all languages, a sort of fundamental Universal Grammar.

For Husserl, as well, a proposition is not a simple string of words, but a signification structure whose meaning is a function of the meanings of its constituents insofar as they appear to belong to specific semantic categories. Therefore, certain *connection semantic rules* are necessary to integrate incomplete parts with the right parts in order to semantically complete them (similarly to Frege's notion of saturation of a function by its arguments). There are categories that form a basis for applying the elements of other categories as operators, the result of this application becoming a new base for further applications, as he describes when he establishes the laws of the compounding of meanings and the pure logico-grammatical theory of forms: "all possible forms of concrete formations are in systematic dependence on a small number of primitive forms, fixed by existential laws; and these forms can therefore be extracted by pure construction" (*Fourth Logical Investigation*, §13). This means that for Husserl, as well as for Frege, the signification of the linguistic expressions is not limited to referential meaning. Lexical items only acquire meaning in the composition with other elements of certain categories. Therefore, it is the compounded significance, the sentence, the

[4]Although this principle is traditionally attributed to Frege, was not he who formulated it, but Rudolf Carnap, who ascribed him the following principles of interchangeability: "First principle [...] the *nominatum* of the whole expression is a function of the *nominata* of the names occurring in it. [...] Second principle [...] the sense of the whole expression is a function of the senses of the names occurring in it." (Carnap 1947: 121). Later, Donald Davidson spread the idea that the Principle of Compositionality is due to the distinction between meaning and reference in Frege: "If we want a theory that gives us the meaning (as distinct from reference) of each sentence, we must start with the meaning (as distinct from reference) of the parts. Up to now we have been following Frege's footsteps; thanks to him the path is well known and even well worn." (Davidson 1967: 306).

[5]This is the real Frege's Principle, which was set by him in *The Foundations of Arithmetic*: "Mann muss die Wörter im Sätze betrachten, wenn man nach ihrer Bedeutung fragt [...] Es genügt, wenn der Satz als Ganzes einen Sinn hat; dadurch erhalten auch seine Teile ihren Inhalt." (Frege 1884: secc. 60).

basic unit of meaning, not the words. Thus, the relations of the parts to the whole, as they were originally proposed by his teacher Brentano, and presented in the *Third Logical Investigation* as a kind of mereotopology applicable to concepts, is not only the basis for understanding concepts, but also the meaning of linguistic categories. This seems to be the source of the subsequent distinction between basic and functorial categories established some years later by the Polish School, as it will be seen.[6]

It is evident, therefore, that we can already find in Husserl's proposal that the mode of composition depends on the chosen set of categories of significance, observing a certain set of universal principles, as they were described by Casadio (1988):

1. Any linguistic expression must belong to a category of significance.
2. Any meaningful expression is the result of the integration of its parts, depending the integration mode on the categories of significance to which each part belongs.
3. By replacing a part of a meaningful expression by an expression of a different category of significance, the first ever becomes non-meaningful.

16.3 Algebraic Categorial Grammar

16.3.1 The Polish School

As said above, Husserl proposed a rule-based grammar over semantic connections through which it was possible to integrate the meaningful incomplete parts with the right parts to complete them, in a similar sense to the saturation of a function by its arguments. Categories are then the elements to be combined to get new complex categories like functions that apply over other functions. This is the idea developed by the Polish School of the Lwów-Warsaw Circle.

The first formal applications of Husserl's proposal treat complex categories like functions that formalize predication and other syntactic connections. The works by Lesniewski (1927–1931) and Ajdukiewicz (1935) are the most representative in this sense.

Lesniewski's Grammar of Semantic Categories (*Semantische Kategorien*) tries to replace with advantage Russell's Theory of Types. His system consists of three interdependent and nested axiomatic theories: protothetic, ontology and mereology.

[6]Frege's idea of replacing the subject-predicate structure by the function-argument structure as a representation of the enunciative sentence was fundamental to the development of the theory of types and Russell's proposal of a hierarchy of types to solve the so-called *Frege's paradox*. Russell's theory of types influenced Lesniewski's grammar of semantic categories, but it was more appropriate for formal languages than for natural languages. However, it was Husserl's theory of a pure grammar which served as a model for Ajdukiewicz's notion of a logical syntax of natural language because of its more natural conception of semantic categories.

Unlike Frege—for whom the domains were objects and functions of various levels—and Russell—who postulated a hierarchy of propositional functions—, Lesniewski proposes categories referring to classes of expressions instead of classes of entities, assuming an essentially nominalistic ontology, based on significations rather than on entities. We consider this change conceptualistic versus the physicalist realism that is behind Set Theory and Model Theory, subsequently developed by his disciple Alfred Tarski.

But Lesniewski never explicitly formulated a theory of semantic categories applicable to natural language. This task was performed by Kazimierz Ajdukiewicz. When Ajdukiewicz (1935) reformulates the idea of a categorial grammar, he is more interested in natural language than Lesniewski. His proposal of a categorial grammar is based on two basic categories (n, s)—the same basic categories of Lesniewski—which lead to complex categories through a defined relation represented by Ajdukiewicz like a mathematical ratio $\frac{A}{B}$, where A and B are any two simple or complex categories. Every expression of a language, simple or complex, belongs to a basic category or to a functional type defined over the two basic categories. The categorial grammar so obtained is known as *AB Categorial Grammar* (after Ajdukiewicz (1935) and Bar-Hillel (1950, 1953), who reintroduced it in the logical debate about the mathematical structure of the grammar in the 1950s). To avoid the complexity arising from the vertical representation of complex categories in Ajdukiewicz, Lambek (1958) introduced a notation using horizontal directional functors \ and /. Following this notation, we can briefly define AB Categorial Grammar as follows:

1. Category n is the category of those expressions that refer to an individual.
2. Category s is the category of those expresions that refer to a proposition.
3. The function type $A\backslash B$ is interpreted as the type of an expresion that results of type B when it is preceded by an expression of type A.
4. The function type B/A is interpreted as the type of an expresion that results of type B when it is followed by an expression of type A.

16.3.2 An Algebraic Grammatical Calculus

By mid-century, Joachim Lambek developed a categorial type system to treat the combinatorial possibilities of the syntax of natural language from a computational perspective. On the basis of Lambek's proposals (1958), we can reformulate in a more contemporary style the definition of a categorial grammar as a kind of algebraic calculus in the following terms:

- *Categorial Grammar* is a tuple $\langle \Sigma, Prim, Tp, \rhd \rangle$, where:
  - Σ is a finite set of symbols
  - $Prim$ is the set of primitive types
  - $Tp(Prim)$ is the set of all types built over the set of primitive types such that it is the smallest set that satisfies $Prim \subseteq Tp(Prim)$ and if $X, Y \in Tp(Prim)$ then $(X/Y), (Y\backslash X) \in Tp(Prim)$

– $\rhd$ is a relation that assigns a lexical element to a categorial type such that $(\rhd) \subseteq Tp(Prim) \times \Sigma$

For example, consider the structure of this simple sentence in English:

The	experienced	grammarian		calculates	this	sentence
n/n	n/n	n		(n\s)/n	n/n	n
	n				n	
	n				n\s	
		s				

The parser is ((the (experienced grammarian)) (calculates (this sentence))), corresponding to the following phrase structure:

$[_S\ [_{NP}[_{Det}\textit{the}[_{Adj}\textit{experienced}[_N\textit{grammarian}]\ [_{VP}[_V\textit{calculates}[_{NP}[_{Det}\textit{this}[_N\textit{sentence}]\ ]$

We can detect several characteristics of categorial grammars in this example:

1. Every word is assigned a category or a categorial type.
2. Complex concepts as "experienced grammarian" or "this sentence" are assigned a categorial type calculated as a function over the types of the simplest ones.
3. All the process is a kind of predication: we predicate "experienced" of "grammarian", "the" of "experienced grammarian", "calculates" of "this sentence", and "calculates this sentence" of "the experienced grammarian", thereby obtaining more complex significations in each level of rules application.

This functional calculus is both a combinatorial logic for syntax and a predicate function over meaningful expressions, so that the concept involved in the meaning of "grammarian" is not the same than the concept involved in the complex expression "experienced grammarian", where the adjective modifies the extension of the nominal reference. Similarly, when using the determiner "the", we are selecting part of the general meaning of the complex expression "experienced grammarian", which implies a difference in the reference and a distinction on the concept.[7]

16.3.3 From Syntax to Meaning

At this point, we can easily expand to semantics. As defined by Ajdukiewicz and Lambek, Categorial Grammar is a syntactic calculus, but it can be expanded to semantics, reinterpreting categories and functional types. That was the intention of Richard Montague when he proposed the semantic theory that bears his name (Montague 1973/1974).

[7] In linguistics, the function of adjectives in a Noun Phrase can be interpreted as a restriction of the reference of the nouns they modify, whereas determiners select a subset of the reference of nouns.

Montague Semantics is based on the Categorial Grammar proposal, but it is semantically oriented:

- Two basic categories: e and t—correspondig to the AB Categorial Grammar categories n and s—such that e is the category for linguistic objects that refer to entities (for instance, proper names or personal pronouns) and t is the category for linguistic objects that can be assigned a truth value (for instance, sentences).
- Just one function to get categorial types, represented by ordered tuples: $\langle e, t\rangle$ | $\langle e, \langle e, t\rangle\rangle$ | $\langle\langle e, t\rangle, t\rangle$ | $\langle t, t\rangle$ | $\langle\langle e, t\rangle, \langle e, t\rangle\rangle$ | $\langle e, \langle\langle e, t\rangle, \langle e, t\rangle\rangle\rangle$

Montague's approach is truth conditional—the meaning of sentences is given by specifying their truth conditions—, model theoretic—i.e.: it includes the construction of abstract mathematical models of external references that constitute the semantic values of the expressions in the object language—and makes use of the notion of possible worlds (Dowty et al. 1981: 4–13). It also makes use of a higher-order type-theoretic language to represent an infinite number of interpreted syntactic categories.

It is possible to combine Montague Semantics with a set of rules based on Lambek Sequent Calculus, what turns it into a grammar—we will call it Montague Grammar—capable of generating well-formed complex linguistic expressions from simpler expressions. For instance, we may consider these simplified rules and the corresponding motivations to accept them in a calculus of categorial types:

Application: This is the only rule in AB Categorial Grammar and the simplest one. We can define it as a unidirectional (Ajdukiewicz) or as a bidirectional rule (Bar-Hillel, Lambek), but in a basic Montague Grammar it is just the rule for getting complex modified expressions by simple predication:

$$\frac{A \qquad \langle A, B\rangle}{B}$$

Composition: This rule allows a pair of functions that share a type to be "composed" in a new single function where its value is that of the functor function and its argument is that of the argument function:

$$\frac{\langle A, B\rangle \qquad \langle B, C\rangle}{\langle A, C\rangle}$$

Raising: This rule states that an expression of a single type (say a name) may be raised to a functional categorial type (say a noun phrase):

$$\frac{A}{\langle\langle A, B\rangle, B\rangle}$$

Division: The intuition behind this rule is that "every sentence-modifying adverb is also a predicate-modifying adverb, symbolically, $s\backslash s \rightarrow (n\backslash s)\backslash(n\backslash s)$" (Lambek 1958 [1988]: 165):

$$\frac{\langle B, C\rangle}{\langle\langle A, B\rangle, \langle A, C\rangle\rangle}$$

It can be proved that Montague Semantics supplemented with the rules of Lambek Calculus results in a grammar that keeps the generative capacity of a context free Generalized Phrase Structure Grammar (Pentus 1997/1996). In the next examples, a few rules of GPSG are redifined in terms of Montague's categorial types and "calculated" by applying some of the rules above:

$S \rightarrow NP\ VP$: $t \longrightarrow \frac{\langle\langle e,t\rangle,t\rangle \quad \langle e,t\rangle}{t}$ Appl.
$NP \rightarrow N$: $\langle\langle e, t\rangle, t\rangle \longrightarrow \frac{e}{\langle\langle e,t\rangle,t\rangle}$ Rais.
$NP \rightarrow Det\ CN$: $\langle\langle e, t\rangle, t\rangle \longrightarrow \frac{\langle\langle e,t\rangle,\langle\langle e,t\rangle,t\rangle\rangle \quad \langle e,t\rangle}{\langle\langle e,t\rangle,t\rangle}$ Appl.
$VP \rightarrow TV\ NP$: $\langle e, t\rangle \longrightarrow \frac{\langle e,\langle e,t\rangle\rangle \quad \langle\langle e,t\rangle,t\rangle}{\langle e,t\rangle}$ Comp.

The last one, for example, is a rule that generates a Verb Phrase structure (*VP*) from a transitive verbal head (*TV*) and a Noun Phrase (*NP*). *VP* is a linguistic object of categorial type $\langle e, t\rangle$—i.e.: a linguistic object that needs an object of category e to become an object of category t. On the other hand, a transitive verb (*TV*) is a linguistic object that needs two objects of category e (a subject and a direct object) to become an object of category t. Finally, *NP* is a linguistic object that needs a *VP* to become an object of category t. It is easy to see that the composition of *TV* and *NP* is a complex expession of categorial type $\langle e, t\rangle$.[8] We can illustrate this correpondence with the top-bottom/bottom-top syntactic tree in Fig. 16.1.

16.3.4 Categories, Unification and Argument Structure

The most important objection to categorial grammars is that they are equivalent to context-free phrase structure grammars, which makes them not suitable for the description of certain syntactic phenomena. Nevertheless, these grammars are useful for the semantic description of phenomena such as quantifier scope, anaphoric relations, ambiguities between *de dicto* and *de re* interpretations of certain terms or the distinction between extensional and intensional verbs like propositional attitudes.

But the very important fact is that categorial grammars—AB Categorial Grammar, Montague Grammar, etc.—put together syntactic combinatorial rules and semantic categories, developing the old idea that linguistic expressions, simple or complex, are combined to give rise to new meanings, as the basis of our infinite ability to generate concepts, by means of logical mechanisms based on predication and functional composition.

[8]The correspondence between sentence constituents in a phrase structure grammar and the categorial types proposed by Montague can be easily established from the syntactic functions and the generic meaning attributed to the different parts of speech (Montague 1973/1974: 249–250).

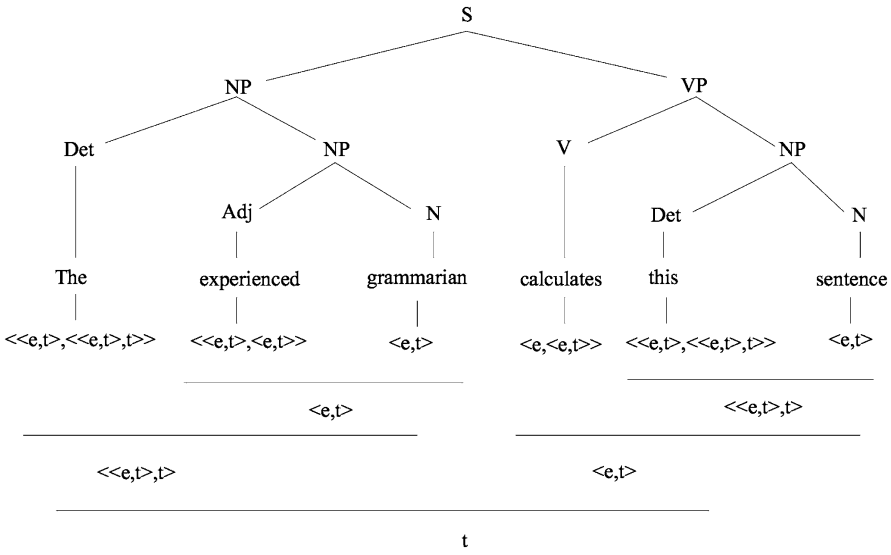


Fig. 16.1 Top-bottom/bottom-top syntactic tree

Moreover it is possible to combine Montague Grammar with a Dynamic Predicate Logic (Groenendijk and Stokhof 1989, 1991) as well as to use Categorial Grammar as a basis for a Unification Grammar based on types (Uszkoreit 1986, Pollard and Sag 1994).

Unification Grammars extend previous representation languages insofar as they include a clear denotational semantics and allow the encoding of grammatical knowledge independently of any specific processing algorithm. As Sikkel and Nijholt say:

> Unification grammars treat syntactic and semantic information in a uniform manner. One can reduce the role of syntax and consider syntactic category as a feature like any other (...) [T]he efficiency of unification grammar parsing can be increased by retrieving an (implicit) context-free backbone from a unification grammar that covers more than just the CAT feature and using this context-free part for syntactic analysis. (Sikkel and Nijholt 1997: 96)

This means that we can define the basic categories and types of a natural language grammar as feature structures; for instance, basic category *n*. Let this be the feature structure for common nouns:

$$noun \equiv \begin{bmatrix} CAT-TYPE & \langle e,t\rangle \\ CONT & +,- \\ DEF & +,- \\ CASE & NOM, ACC, OBL \\ LEVEL & 1,2 \\ ROLE & AG, PAT, GOAL, INSTR, DEST, LOC \\ AGR & \begin{bmatrix} NUM & SING, PL \\ GEN & MASC, FEM, NEUT \\ PERS & 3rd \end{bmatrix} \end{bmatrix}$$

This structure contains all the information necessary to combine a linguistic object of category *n* with other linguistic objects, according to categorial combination rules, but limiting the possible complex expressions that can be obtained to those in which the unification of features matches.

This is, of course, just a single way among many others to represent the semantic and syntactic content of a linguistic object, using unification-based feature structures (Uszkoreit 1986, Pollard and Sag 1994, Villavicencio 2002, Cooper 2008). The important point here is that the treatment of categories as feature structures opens a door to apply the rules of a categorial calculus on a basic argument structure of the sentence with feature unification.

The argument structure of a sentence is a basic predicate structure that supports recursivity (i.e.: all its arguments can be replaced as well by complete argument structures):

$$MOD((PRED(arg_2), (arg_3))arg_1)SAT$$

In an argument structure, only the predicate *PRED* and the external argument arg_1 are necessary. The inner arguments arg_2 and arg_3 are not necessary—they depend on the kind of predicate—, *MOD* is any kind of modality and *SAT* might be a complementary squence of terms or structures. All the arguments in the structure help complete the meaning of the predicate by playing a thematic role ($\theta - role$) like *agent, patient, goal, instrument, experiencer, beneficiary, source, location*...

Argument structures are the basic framework in which meanings are combined to give rise to complex grammatical structures according to the grammar of each language. So to speak, an argument structure defines the functional and predicative universal relations that allow to interpret any sentence in terms of its grammatical structure and the cognitive contents involved in the lexicon.[9]

So, for Derek Bickerton, the argument structure is the basis of syntax. The concatenation of signs is not enough to get syntax, as can be seen in the early stages

[9]The relationship between predicates and arguments depends on the valence of the latter. It was first expressed by the French linguist Lucien Tesnière, for whom an argument is an expression that helps complete the meaning of the predicate (Tesnière 1959[1965]: 128).

of language development in children. Basic predication leads to the formation of phrases (noun, verbal, prepositional phrases) that end up being used by the child to form increasingly elaborate clauses from a grammatical point of view. These clauses reflect the argument structure of the sentence and are the foundation of syntax:

> Before there was syntax, there was only semantics. So, if you are looking for the very first stages in the development of syntax, you have to look in semantics for whatever is the most syntaxlike thing. Argument structure is the most plausible candidate. It involves meaning (the meanings of the thematic roles, agent and so on, and their relation to the verb meaning) but it can be readily mapped onto linguistic output to provide that output with structure (...). (Calvin and Bickerton 2000: 50)

Even more, if we believe the neurologist William Calvin, nouns, adjectives or verbs are stored in our memory in different brain locations (Calvin and Bickerton 2000: 58–61). Only children's learning processes and the acquisition of syntax as a process of combination of categories and meanings explain how so different mental objects can be combined to express complex concepts within the predicate argument structure. No doubt this is one of the most fruitful path open nowadays to be explored by linguistics.

16.4 Conclusion

We can trace up to the ancient Greece the notion of predication as the fundamental function to assign meaning to complex expressions. This notion is mainly due to the Aristotelian distinction between subject and predicate that underlies the distribution of the lexicon in semantic-conceptual categories. Word meaning so conceived is in harmony with some attached ontology and epistemology, resulting in universalist and rationalist theories about the role that represents language in perception, comprehension and conceptualization. With the emergence of phenomenology and formal models of logic in the twentieth century, such theories have led to the functionalist proposals that have been known as Categorial Grammars.

Categorial Grammars are a set of grammatical formalisms in which categorial types are conceived as functions that relate a (complex) meaning and certain combinatorial properties to get more complex categories. These formalisms can be seen as a way of representing concepts as the result of different levels of predication among lexical items and generalized categories. Categorial types are the objects on which combination rules are applied in a categorial grammar, and not the words themselves. Different kinds of predicative relations can be reduced to just a few categories and a few rules for getting categorial types, so that each category and each categorial type provide the basic semantic and syntactic information of the linguistic objects that are assigned, and a calculus on syntactic categories and types can be defined, in a certain way, as an interpreted calculus.

Of course, simple and complex categories must be translated into linguistic expressions that fit specific morphological and syntactic features of a particular natural language. Though the most important aspect for combining them is—from

the point of view of the concept—their meaning, morphosyntactic aspects should be included in any calculus of categorial types. To do this, categories and categorial types can be treated as feature structures that participate in recursive argument structures built over predicates and arguments. This proposal is certainly very elegant, because it helps define a syntactic interpreted calculus and link it to certain evolutionary theories about the emergence of syntax from semantics as well as to the acquisition process of language by children.

Acknowledgements This paper is part of the research project "Awareness, Logic and Computation", financed by the Government of Spain. I would also like to acknowledge all the help and support of the members of the Research Group on Logic, Language and Information at the University of Seville (GILLIUS), as well as the suggestions of an anonymous referee that have been very helpful to improve the original text.

References

Ajdukiewicz, K.: Die syntaktische Konnexität. Studia Philosophica **1**, 1–27 (1935)
Bar-Hillel, Y.: On syntactical categories. J. Symb. Log. **15**, 1–16 (1950)
Bar-Hillel, Y.: A quasi-arithmetical notation for syntactic description. Language **XXIX**, 47–58 (1953)
Calvin, W.H., Bickerton, D.: Lingua ex Machina. Reconciling Darwin and Chomsky with the Human Brain. MIT, Cambridge (2000)
Carnap, R.: Meaning and Necessity. Chicago University Press, Chicago (1947)
Casadio, C.: Semantic categories and the development of categorial grammars. In: Oehrle, R., Bach, E., Wheeler, D. (eds.) Categorial Grammars and Natural Language Structures, pp. 125–151. Reidel, Dordrecht (1988)
Cooper, R.: Type theory with records and unification-based grammar. In: Hamm, F., Kepser, S. (eds.) Logics for Linguistic Structures, pp. 9–34. Mouton de Gruyter, Berlin (2008)
Davidson, D.: Truth and meaning. Synthese **17**, 304–323 (1967)
Dowty, D.R., Wall, R.E., Peters, S.: Introduction to Montague semantics. Reidel, Dordrecht (1981)
Frege, G.: Grundlagen der Arithmetik. Eine logisch mathematische Untersuchung über den Begriff der Zahl. Wilhelm Koebner, Breslau (1884)
Groenendijk, J., Stokhof, M.: Dynamic montague grammar. In: Papers from the Second Symposium on Logic and Language, pp. 3–48. Akademiai Kiadoo (1989)
Groenendijk, J., Stokhof, M.: Dynamic predicate logic. Linguist. Philos. **14**, 39–100 (1991)
Husserl, E.: Logische Untersuchungen. Niemeyer, Halle (1990–1901)
Kanerva, P.: Sparse Distributed Memory. MIT, Cambridge (1988)
Lambek, J.: The mathematics of sentence structure. Am. Math. Mon. **65**(3), 154–170 (1958). Reprinted in Buszkowski, W., Marciszewski, W., van Benthem, J. (eds.) Categorial Grammar, pp. 153–172. John Benjamins, Amsterdam (1988)
Lesniewski, S.: O podstawach matematyki. Przeglad Filozoficzny **30**, 164–206 (1927–1931); **31**, 261–291 (1928); **32**, 60–101 (1929); **33**, 77–105 (1930); **34**, 142–170 (1931)
Montague, R.: The proper treatment of quantification in ordinary english. In: Thomason, R. (ed.) Formal Philosophy: Selected Papers of Richard Montague, pp. 247–270. Yale University Press, New Haven (1973/1974)
Pentus, M.: Lambek Calculus and Formal Grammars, American Mathematical Society. (Transl. from the original Ph.D. Dissertation, Moscow State University (1997/1996))

Pollard, C., Sag, I.: Head-driven Phrase Structure Grammar. Center for the Study of Language and Information (CSLI) Lecture Notes. Stanford University Press/University of Chicago Press, Stanford (1994)

Robins, R.H.: A short History of Linguistics. Longman, London (1969)

Sigman, M., Cecchi, G.A.: Global organization of the wordnet lexicon. PNAS **99**(3), 1742–1747 (2002)

Sikkel, K., Nijholt, A.: Parsing of context-free languages. In: Rozenberg, G., Salomaa, A. (eds.) The Handbook of Formal Languages, vol. II, pp. 61–100. Springer, Berlin (1997)

Tesnière, L.: Éléments de syntaxe structurale. Klincksieck, Paris (1959); 2nd edition (1965)

Thorndike, E.L., Woodworth, R.S.: The influence of improvement in one mental function upon the efficiency of other functions. Psychol. Rev. **8**, 247–261, 384–395, 553–564 (1901)

Uszkoreit, H.: Categorial unification grammars. In: Proceedings of the 11th International Conference on Computational Linguistics. ACL, Bonn (1986)

Villavicencio, A.: The acquisition of a unification-based generalised categorial grammar. Technical report 533, UCAM Computer Laboratory, Cambridge (2002)

Whorf, B.L.: Language, Thought and Reality. Selected Writings of Benjamin Lee Whorf. MIT, Cambridge (1956)

Part IV
A Critical Interlude

Chapter 17
On Leonard Nelson's Criticism of Epistemology

Jan Woleński

Abstract This paper analyses proofs of impossibility of epistemology formulated by Leonard Nelson. He proposed two such demonstrations. The first proof tries to show that no criterion of knowledge is possible. Nelson's second argument considers the sentence *B* '*A* is a piece of knowledge' as being synthetic (in Kantian sense). On the other hand, Epistemology cannot employ problematic premises. Hence, it consists of analytic sentences. Now, epistemology is impossible because synthetic sentence cannot be derived from purely analytic ones. The analysis conducted in the papers intends to locate Nelson's arguments in contemporary discussions about the status of epistemology. In particular, it is argued that the second proof deserves attention of contemporary epistemologists.

Keywords Neo-Kantianism • Logic • Critical method • Fries trilemma • Justification • Petitio principi • Knowledge

Leonard Nelson is almost completely forgotten by contemporary English speaking philosophers. Since 1967 (the entry "Nelson, Leonard" by G. Henry–Hermann, in *The Encyclopedia of Philosophy*, ed. by P. Edwards, vol. 5, Collier Macmillan, London) his name has not appeared in encyclopedias (including *The Routledge Encyclopedia of Philosophy*, ed. by E. Craig, Routledge, London 1998 and *Stanford Encyclopedia of Philosophy*, the latter until now) or philosophical handbooks and companions. Only three (as far as I know) Nelson's book were translated into English, namely *Socratic Method and Critical Philosophy* (Yale University Press, New Haven 1949; see References at the end of this paper), *System of Ethics* (Yale University Press, New Haven 1956) and *Progress and Regress in Philosophy*, vols. I–II (Basil Blackwell, Oxford 1970–1971). Yet, Nelson's importance in the history of contemporary philosophical thought cannot be denied. He established the so-called Neo-Friesian School, frequently considered as the third major Neo-Kantian center, besides the Mahrburg School (Hermann Cohen, Paul Natorp) and the

J. Woleński (✉)
Department of Social Sciences, University of Informatics, Management and Technology, Rzeszów, Poland
e-mail: wolenski@if.uj.edu.pl

J. Redmond et al. (eds.), *Epistemology, Knowledge and the Impact of Interaction*, Logic, Epistemology, and the Unity of Science 38, DOI 10.1007/978-3-319-26506-3_17

Badenian School (Wilhelm Windelband, Heinrich Rickert). Nelson was very active and productive in almost all domains of philosophy, particularly in epistemology, ethics, philosophy of law, philosophy of education, but also in mathematical logic (important papers on logical antinomies; Paul Bernays and Kurt Grelling, distinguished logicians, and Gerhard Hessenberg – one of the first champions of set theory belonged to the Neo-Friesian School; Nelson himself graduated in mathematics). Nelson's philosophical ideas could have influenced David Hilbert. I fact, looking at the Hilbert program in the foundations of mathematics through the glasses of Nelson's critical philosophy constitutes one of interpretations of Hilbert's finitism (see Peckhaus 1990; I will return to this question below). Another possible path of Nelson's influence brings us to Karl Popper and his fallibilism. The Popperian methodological analysis of science begins from recalling the so-called Fries trilemma (see below) popularized by Nelson at the beginning of the twentieth century.

This paper examines Nelson's two arguments against the possibility of epistemology. My task exceeds purely historical interests. I would like to embed Nelson's attitude toward epistemology into a more contemporary perspective and show how it might illuminate present controversies concerning the nature of epistemology and the concept of knowledge together with the problem of epistemic justification. Moreover, I would like to establish the scope of both proofs in the sense of pointing out which kinds of epistemology are targeted by Nelson's objections. I begin with general remarks about some philosophical views of Nelson and their background. Nelson followed Jacob Friedrich Fries, one of Kant's followers. Fries tried to interpret Kantianism without any commitment to transcendentalism, because, as he argued, we cannot prove the objective validity of knowledge via transcendental elements. He qualified his reading of Kant as anthropological. According to Fries, any justification of the objectivity of knowledge by appealing to a priori structure of our thinking inevitably and fatally falls into an unsolvable trilemma with dogmatism (assuming some propositions without justification), vicious circle (justifying propositions by statements grounded by the former) and regressum ad infinitum (admitting the infinite chain of propositions to be justified) as its horns. More specifically, Kant's transcendental deduction of categories suffers from the fallacy of the trilemma, because it cannot be developed without appealing to dogmatism, circularity in justification or proceeding by regress to infinity. Fries proposed the regressive method (see below) as an indispensable device for discovering ultimate presuppositions of knowledge. Passing from direct evident acquaintance to indirect knowledge became the main problem in this approach. Fries tried to solve the problem of knowledge by arguing that we possess the direct non-evident knowledge, which does not arise from inductive inference. Doubtless, it was a novelty, because the traditional view attributed directness exclusively to evident knowledge. This position was considered by Fries as the essence of the anthropological critique (in Kant's sense) of pure reason. Most commentators of Fries' solution of the problem of knowledge acquisition agree that the boundary between his anthropologism and psychologism cannot be sharply drawn. This constitutes the main objection against Fries' epistemology as unclear in its very foundations.

Nelson inherited the fundamental points of Fries' philosophy. However, Nelson felt that the methodological level of his master required a further elaboration as not sufficiently sophisticated in the light of new developments in philosophy and logic. Nelson fully agreed that the regressive method has to be employed in philosophy and that it differs from induction. The regression should also be distinguished from deduction associated with the axiomatic method and having internal limitations, because it leaves axioms without justification. Thus, stopping at deduction does not only liberate us from the Fries trilemma. In particular, the limitation of the justification strategy to deduction (the progressive method as Nelson termed it) would result in the involvement into dogmatism. This means that the basic problem of the regressive method consists in the way in which axioms or other general principles of science can be proved to be valid. The word 'proof' has a wider sense in this context than it possesses in the theory of deduction. For Nelson, deductive proofs apply to indirect propositions derived from axioms. How to prove axioms? Nelson, following Fries, assumed that our reason possesses its own resources to proceed by abstraction and to demonstrate axioms as valid (universally true, *gültig* in German) directly not-evident propositions. Although Fries and Nelson expressed several doubts about the transcendental deduction in Kant's sense, regression plays a similar role in their philosophy as the transcendental reasoning in Kantian critique of the pure reason. Unfortunately, this circumstance causes an ambiguity of such crucial terms as 'proof', 'justification' or 'deduction'. For instance, one must be very careful in deciding whether proofs or demonstrations are understood by Nelson metamathematically, that is, as based on the notion of logical consequence or serve as instruments of grounding axioms or other assumptions as being correct.

The Neo-Friesians attempted to employ the regressive method in the philosophy of mathematics. The idea of so-called critical mathematics (see Peckhaus 1990, pp. 123–168) well illustrates the essence of regressive method and its functioning in a concrete case. Mathematical axioms obtain their justification by demonstrating their consistency (the main criterion), completeness, mutual independence, and possible applications in science. Clearly, the project of critical mathematics was influenced by the discussion on the foundations of mathematics at the beginning of the twentieth century, particularly in Germany. In fact, Nelson and other Neo-Friesians considered Hilbert as a typical representative of critical mathematics. Although this qualification of Hilbert's philosophical view about the essence of mathematics should be taken *cum grano salis* as too simplified and neglecting several important features of formalism as a position in the foundations of mathematics, one might certainly forward some historical and substantial arguments for considering Hilbert as somehow related to Nelson and his group. Firstly, the latter belonged to the Göttingen Circle and was highly appreciated by the former. In fact, Hilbert supported Nelson in his unsuccessful attempts to become the ordinarius in philosophy at the University of Göttingen. Secondly and more importantly, we can easily find a textual evidence that Hilbert considered the regressive method as important in mathematical practice, especially for justifying axioms and other general principles of science (see Hilbert 1992, pp. 15–19; this book is based on Hilbert's course on problems of mathematical and scientific problems delivered

in Göttingen in the academic year 1919–1920; Nelson is mentioned on p. 18 as a person who actually introduced the regressive method to mathematics).

Now let us consider the following interpretation of the Hilbert program. We need to build the reliable basis for mathematics. The first step consists in reducing all mathematical concepts to a few simple und simple unproblematic ideas, easily settled by mathematical experience. This task can be achieved by arithmetization of the entire mathematics, because arithmetic suffices as the basis of all remaining mathematical fields. The next stage consists in showing that set theory constitutes the indispensable basis for constructing the conceptual apparatus for arithmetic. Since this very attractive and promising strategy becomes problematic after discovering antinomies of set theory, the convincing consistency proof of arithmetic together with set theory appears as the necessary step in reliable grounding of mathematics. How to define reliability in the considered case? Clearly, the required proof of consistency should be finitary (roughly speaking, realizable in the finite number of steps), because this property guarantees its control by firm evidence, almost similar to empirical evidence. Thus, a good and convincing consistency proof should be realized by purely finitary methods. Clearly, such a description of Hilbert's strategy does not require any appeal to the combination of directness and evidence or even the regressive method. On the other hand, one might be tempted to say that since the finitary consistency proof as the completely reliable grounding of mathematics does not proceed axiomatically, this way gives an almost perfect piece of critical mathematical thinking in Nelson's understanding. Once again, I do not claim that this reading of the Hilbert program is the only possible or even the best one. I only say that the Neo-Friesian could somehow faithfully invoke it as confirming his or her general philosophical insights. Anyway, the regressive method as applied to the problem of consistency sufficiently shows that the method in question requires quite complex reasoning with many patterns and deliberations. Yet, a very serious metaphilosophical issue consists in the question of how far the above mathematical example can be generalized. If we take the morals coming from the entire history of philosophy seriously, we should be very modest in saying that the regressive method provides a new philosophical stone/landmark? Nelson himself was fully convinced that this method will successfully transform philosophy into one of the mature sciences. Thus, he ascribed a fairly revolutionary role to his own philosophical methodology.

Nelson offered two proofs (they are stylized as forwarded deductively) which were supposed to demonstrate that epistemology is impossible (see Nelson 1908, pp. 441–517, Nelson 1911; the second work reproduces Nelson's talk at the 4th International Philosophical Congress held in Bologna in 1911). He considered them as typical applications of the regressive method in philosophy. Both Nelson's arguments assume the following general premise (note that Nelson uses the method of proving by cases):

(*) The fundamental task of epistemology consists in demonstrating objective truth or validity of human knowledge (this assumption simply displays the main problem of epistemology);

(α) The first proof

1. A solvability of the problem stated in (*) requires that we have a criterion, which, when applied to results of our cognition, could decide whether these results are true or not. This criterion (I will refer to it by the letter **C**) the epistemological criterion;
2. **C** is either knowledge or not;

   (a) Assume that **C** is knowledge;
   (a1) If **C** is knowledge, it belongs to the domain of what is just problematic (Nelson assumes that that a piece of cognition is problematic before checking it by **C**);
   (a2) However, **C** is not knowledge, it is problematic only;
   (a3) Contradiction (a) – (a2);
   (b) Assume that **C** is not knowledge;
   (b1) If **C** is to be successfully applied, it must be known as suitable to perform its role as the standard of knowledge;
   (b2) If (b1), **C** should be knowledge;
   (b3) Contradiction (b) – (b2);

3. Since we get a contradiction in every case listed in (2) and because (2) depicts the complete and exhaustive list of possibilities, the problem of epistemology has no satisfactory solution. This just means that epistemology is impossible.

For instance, assume that the epistemological evidence of the criterion **C** consists in evidence (*A* is a piece of knowledge if and only if *A* is evidently true; a more precise definition of evidence is not relevant here). In other words, we need to know that if an act of cognition satisfies **C**, it is knowledge, that is, it manifests itself as evidently true. Consequently, according to our assumption, we infer that **C** satisfies **C.** However, in order to apply our assumption, this step must presuppose that **C** itself is evident. However, this presupposition does not constitute knowledge, because it must be examined by **C**. Thus, we get a contradiction**:** C is knowledge and **C** is not knowledge. We can similarly examine other concrete cases of **C**, for instance, coherence or consensus.

(β) The second proof

1. Since knowledge is something problematic for epistemology, any attempt to solve this question must abstain from accepting something as knowledge;
2. If (1), then epistemology must begin with an analysis of concepts and consequences derived from such an analysis;
3. If (2), epistemology formulates analytic propositions only. Nelson considers the division of propositions into analytic and synthetic as legitimate; the former are defined as obtainable exclusively from concepts (Nelson says that this account reproduces Kant's ideas);
4. Propositions expressed by analytic sentences do not provide new knowledge;

5. If the proposition A is an instance of knowledge, the proposition asserting that the proposition A is an instance of knowledge (for brevity, B = the proposition A is an instance of knowledge) is synthetic, because it is not obtained by the analysis of concepts;
6. If (5), the task of epistemology consists in deriving B from analytic propositions without any use of synthetic ones;
7. Since derivation of synthetic sentences from a set **X** consisting exclusively from analytic sentences is impossible, B cannot be derived in epistemology;
8. If (7), the task of epistemology (= to demonstrate that a given piece of cognition is knowledge) cannot be successfully realized. This just means that epistemology is impossible.

Terminological remarks. (a) Nelson consequently uses the word *Erkenntnis* (knowledge) in original German texts, but, unfortunately, ambiguously. In both Nelson's proofs an instance of knowledge is a cognitive item which satisfies **C**, but, in examples related to (α), beliefs before checking whether they agree with **C** are also qualified as *Erkenntnis*. This ambiguity is preserved in English translation using 'knowledge' and 'cognition; (b) I have already noticed another ambiguity in Nelson, namely that concerning proofs and cognate concepts. I use the terms 'proof', 'deduction' and 'demonstration' in the metamathematical sense. This decision, even if it is somehow at odds with Nelson's intentions, has its legitimacy in the fact that (α) and (β) have the structure of normal deductive proofs.

If one claimed that my choice of terminology misuses Nelson's tasks, I would reply by saying "Well, but my reconstruction intends to show what follows if (α) and (β) are regarded as logical deductions." The term 'justification' has a wide meaning and refers to any procedure employed in showing that a proposition is right, true, correct, etc. In particular, justification can be regression in Nelson's sense. I use utterances 'the statement that A' and 'the propositions that A' as equivalent. This usage does not commit me to recognizing propositions as abstract entities. Propositions are understood here as meanings expressed by sentences. For brevity, the sentence 'proposition that A' is rendered by the sentence 'proposition A'. I could also say 'a proposition expressed by the sentence A', but it would result in passing to the metalanguage in which the sentence A is formulated. Consider the sentence 'Poland became a member of EU in 2004'(this example will be employed below). According to the above explanations, one can say 'the proposition (or the statement) that Poland became a member of EU in 2004 ...' or 'the proposition expressed by the sentence 'Poland became a member of EU in 2004' ...'. The latter belongs to a suitable metalanguage, the former preserves the linguistic level of 'that A'. Note that adding the prefix 'that' cannot be eliminated in concrete examples. The proposed usage tries to avoid semantic ascent in order to use Quine's apt terminology. Self-referential contexts are not dangerous in epistemology, contrary to formal semantics.

One thing should be noticed at once. Nelson's arguments concern the impossibility of epistemology, but they do not say that knowledge cannot be achieved at all. Thus, this very epistemological position must be sharply contrasted with that of skepticism. The skeptical view denies that we can gain knowledge. Consequently, the skeptic argues within epistemology that knowledge is just impossible; I did

not enter here into a very fundamental and frequently discussed question whether skepticism about knowledge can be consistently formulated (= is skeptical epistemology coherent? I omit the discussion of Nelson's argument intended to demonstrate inconsistency of skepticism). In other words, skepticism accepts epistemology, but rejects the possibility of knowledge (see below). Nelson also discusses the problem of knowledge, but his solution cannot be described in details. Roughly speaking, Nelson argues that if we restrict knowledge to something indirect and obtainable by proof, that is, by assuming that every knowledge is inferred from another knowledge, we will inevitably fall in the Fries trilemma. An appeal to direct, perceptual knowledge gives no way out, because it does not solve the question of justification of propositions. This reasoning suggests that the actual possibility of knowledge strongly depends on direct non-evident knowledge, which legitimates the regressive method as sound. Nelson's essential step rejects the identification of knowledge with propositions. I leave this issue without further comments except noticing that contemporary cognitive science gives a support to such a view although it requires making a distinction between knowledge as *episteme* and knowledge as cognition.

Both Nelson's proofs deserve several comments. First of all, the assumption (*) points out how Nelson understands epistemology (*Erkenntnistheorie*, theory of knowledge). His understanding follows Kantian and Neo-Kantian (particularly in the version of the Badenian School) transcendentalism; I will call it the transcendental conception of epistemology. Such a view was prevailing in German academic philosophy at the end of the nineteenth century. According to this view, epistemology acts as the highest court sentencing what belongs to the very scope of knowledge (denote it by **K**) or does not belong to it. If A pretends to be an element of **K**, it must pass the verdict of epistemological trial based on a proof. Yet this claim, usually strengthened by considering epistemology as one of the sciences (an additionally combined with the view that theory of knowledge provides justification for the ultimate presuppositions of scientific research) can be interpreted in two ways:

(i) as a factual assertion (the sentence '$A \in \mathbf{K}$' is provable in epistemology);
(ii) as a postulate (the sentence '$A \in \mathbf{K}$' should be provable in epistemology).

Since Nelson uses the word *prüfen* (to prove) provability must be seriously taken, even if we agree that to 'have a proof' means 'to be justified on the basis of commonly accepted procedures' (a liberal or very intuitive understanding of provability; see also terminological remarks above).

Ad (i) This assertion is obviously false. If A belongs to science or even to ordinary knowledge, it is provable within these domains. Clearly, the borderline between science and philosophy (epistemology in our case) cannot be sharply delineated and we encounter some problematic cases which might be eventually pointed out as examples of using epistemology in proving that something belongs to **K**. Take the Hilbert program once again. Assume that we analyze the equivalence (this is a simplification, because the criterion of mathematical knowledge is more complicated under Hilbert's view):

(#) a theory **T** belongs to mathematical knowledge if and only if **T** is consistent

One can interpret (#) epistemologically and argue that the concept of consistency belongs to the theory of knowledge. For instance, the representatives of critical mathematics could say that we prove consistency by the regressive methods as an epistemological method. However, the problem of consistency became a mathematical issue in hands of Hilbert and his followers. This question obtained a surprising (relatively to Hilbert's expectations about the role of consistency proofs in mathematics as the ultimate criterion of mathematical validity) but exact solution in the Gödel incompleteness theorems. I do not deny that these results have very interesting epistemological consequences and interpretations, but I see no reason to maintain that the validity of mathematical (or any other scientific) knowledge is proved in epistemology, not in mathematics (or science) itself. In other words, epistemological presuppositions and principles do not function as logical premises of scientific theorems or their deductive consequences.

Ad (ii) The postulate that *Erkenntnistheorie* should prove what is knowledge and what is not (= what is valid knowledge and what is not), has a justification inside the transcendental conception of epistemology. In particular, since the Badenian Neo-Kantians considered validity as a normative concept operating in the *Sollen*, but not in the *Sein*, they considered the sentence '$A \in \mathbf{K}$' as decidable inside epistemology. However, nothing substantially more can be invoked in favor of (ii) except this metaphilosophical proposal, which boldly ascribes a position of a super-science to the theory of knowledge (and to philosophy in a more general perspective). Clearly, other programs of epistemology are not damaged by Nelson's first argument at all. Take epistemology in the naturalistic setting (I do not suggest that it is correct). Roughly speaking, naturalism prefers to speak about cognition than knowledge. Hence, epistemology (or cognitology to use a neologism) appears as close to cognitive psychology and sociology and certainly belongs to the domain of (positive) science. And naturalized epistemology has no ambitions to act as a validity-tribunal for particular cases of knowledge. Analytic epistemology (I am a defender of this kind of philosophy, but this point does not matter in the present context) is another fairly instructive case. It tries to analyze various concepts, like knowledge, justification, perception, the object of knowledge, etc. For instance, consider a famous question 'Does $A \in \mathbf{K}$ imply, A is true?'. Even if we agree that the answer belongs to epistemology, this part of philosophy not necessarily pretends to prove that a concrete A is true or not.

Nelson's own treatment of epistemology is not transparent in all respects. When he argues that regarding all instances of knowledge as indirect (= obtainable by proof) has no justification in psychological data, his reasoning looks as produced by a typical naturalist. On the other hand, Nelson declares his almost full faithfulness to Kant's project of epistemology as the correct critique of the pure reason according to well-established rational principles. Another complication stems from his approval of Fries, because (see above) anthropologism looks, perhaps contrary to actual intentions of Nelson and Fries, as a certain compromise between naturalism and transcendentalism. The already mentioned ambiguity of *Erkenntnis* in Nelson's considerations suggests that his ambitious philosophical project was simultaneously tempted by naturalism and transcendentalism. *Nihil novi sub sole,* one could rightly

say. In fact, the entire history of epistemology demonstrably presents itself as heavily burdened by the controversy going back to Parmenides and Plato and concerning the relation between *episteme* and *doxa*. Obviously, Nelson's proofs (α) and (β) go directly against the transcendental conception of epistemology, but they should not be extended beyond this respectable but problematic idea. As far as the issue concerns Nelson's own understanding of epistemology, it seems to be closer to naturalism than to transcendentalism. Not only because he uses psychological data against reducing all knowledge to that proceeding by proofs, but also for his treatment of epistemology as inspiring rather by factual occurrences of error than by focused on how to justify the validity of all possible instances of *Erkenntnis*. Thus, pointing out (see above) a possible link between Nelson's epistemological project and contemporary cognitive science is plausible, but this issue must be left without further comments (see Appendix).

In fact, Nelson argues (see (α) 2b1 below) that in order to make legitimately successful applications the criterion **C** we should know without any doubt that it can function as such (this view was very characteristic for the Badenian School, particularly for Rickert). I cannot find better comments on this claim concerning the epistemological presumption of the criterion **C** than Kazimierz Ajdukiewicz's analysis (see Ajdukiewicz 1949, p. 20–21; page-reference to Eng. tr.) of this issue:

> The skeptics assert that in order to gain justified knowledge it must be arrived by applying a criterion about which we should know beforehand that it is a trustworthy. In other words, in order to gain justified knowledge of any kind we have to have at our disposal according to skeptics not only a trustworthy criterion by means of which we would justify this knowledge but furthermore would have to know that this criterion is itself is trustworthy. It is just here the skeptics' mistake is to be found. The point is that in order to justify an assertion it is sufficient to arrive at it by applying a trustworthy and we do not have to know also that the criterion applied is trustworthy. The knowledge of whether our criterion is trustworthy is not necessary for the justification of the assertion arrived at in accordance with it. It is required only to assure us that we have justified a given assertion. It is one thing to justify an assertion and another to know that one has done so. It is one thing to do something well and it is another to know that one has done so. Thus if the knowledge that the criterion applied in the justification of an assertion is trustworthy is not necessary for the justification of this assertion, then the premise is false from which the skeptics drew their conclusion that the justification of any assertion whatsoever requires an infinite number of steps of reasoning which can never be completed (it is false that it leads to a *regressus ad infinitum*).

Although Ajdukiewicz addressed his remarks to the skeptics, his reasoning also applies to Nelson's argument (in Schlick 1918, p. 90, page-reference is to Eng. translation, we find a short remark pointing out that Nelson confuses to be known and to be an object of knowledge). In fact, here are historical reasons to think that Ajdukiewicz could be inspired by Nelson, because the former studied in Göttingen at the peak of activities of the latter in the same place.

The essence of Ajdukiewicz's argument consists in the statement asserting that knowledge that **C** is trustworthy as applied to a given statement does not constitute a necessary condition for this assertion to be justified (putting this more formally: the statement '*A* is justified' does not entail the statement 'it is known that **C** is trustworthy'). Consequently, it is actually possible that $A \in \mathbf{K}$ and it is not

known that the criterion **C** is trustworthy for *A*. Using a different language (Gilbert Ryle's famous distinction between two different kinds of knowledge), *to know that* does not constitute a necessary condition for *to know how*. Hence, we can know how to correctly operate with **C** without knowing that it is trustworthy (correct, legitimate, valid, etc.) as a (or even the) device of justification in performing our epistemic actions. On the other hand, **C**-operating can require a debate concerning the justification of a given assertion.

Assume that a person *p* claims that an assertion *A* is proved by logic. The proof in question can be problematic or wrong, and a further discussion about the correctness of particular steps is postulated, for instance, during a defense of a PhD in pure mathematics. However, it is difficult to consider such a debate as epistemological in the sense of Nelson's criticism, because it proceeds inside logic applied to a given specific domain. On the other hand, if *A* is correctly inferred, its justification remains proper even if nobody knows that it is such. Consequently, we should distinguish between *knowing how* and *knowing that* in the context of various scientific domains, daily matters and in epistemology as well, disregarding whether the last is a science or not. Provided that Ryle's contrast could be implemented into Nelson's criticism of *Erkenntnistheorie*, we can say that *knowing how* is impossible without *knowing that* just in the traditional epistemological framework. This interpretation helps illuminate Nelson's treatment of the relation between skepticism and his criticism of epistemology. Nelson observes that skeptical rejection of the possibility of knowledge is a consequence of an epistemological prejudice. Now, we can easily identify this prejudice. It consists in the requirement that *knowing that* forms a necessary epistemological condition for *knowing how.*

Nelson's proof (β) is interesting not only as being critical of epistemology, but also as an illustration of the problem how analytic and synthetic sentences are mutually related (in this context I prefer the term 'sentence' to 'proposition', but I return to the latter in next more epistemological paragraphs). Disregarding the proof itself at the moment, let me focus on the latter issue. Although Nelson does not offer a definition of analytic sentences, it is not difficult to prove that consequences of analyticals are analytic as well. However, since we have different kinds of analytic sentences (see Woleński 2004), there is no single argument justifying that if $A \in Cn\mathbf{X}$ (reading: *A* is a logical consequence of the set **X**) and **X** consists exclusively of the analyticals, then *A* is analytic as well (symbolically: $A \in \mathbf{AN}$). Assume that *A* is a tautology of (classical) sentential logic (**SL**). Define analytic sentences as formulas formally provable in **SL**. Thus, $A \in \mathbf{AN}^{\mathbf{SL}}$ if and only if $\vdash^{\mathbf{SL}} A$. Since **SL** is Post-complete (maximally consistent). This means that if we add a non-tautology, let us say *B*, to the stock of theorems of **SL**, the resulting theory $\mathbf{SL} \cup \{B\}$ becomes inconsistent. First-order logic (**FOL**; I consider it as the logic) is not Post-complete. Let $B =$ 'there are exactly *n* objects', where *n* is a natural number > 0. The theory $\mathbf{FOL} \cup \{B\}$ is consistent as satisfied in the semantic model with exactly *n* individuals in its universe. Define $A \in \mathbf{AN}^{\mathbf{FOL}}$ if and only if $\vdash^{\mathbf{FOL}} A$ (**SL**-analyticals are automatically **FOL**-analyticals). Although *B* is not provable in **FOL** and thereby is not analytic, its adding to **FOL** does not produce inconsistency. Finally, consider $B =$ 'a bachelor is an unmarried man' (Quine's famous example). It logically entails

that if a is a female partner of a bachelor, she is not his wife, relatively to the legal meaning of the term 'wife'. In other words, B is provable by pure logic plus the adopted definition of being someone's wife.

Define $A \in \mathbf{AN}^{\mathbf{FOL}\cup\mathbf{DEF}}$ if and only if $\vdash^{\mathbf{FOL}\cup\mathbf{DEF}} A$ (logical tautologies, that is, **SL-**analyticals and **FOL**-analyticals belong to **FOL** ∪ **DEF**-analyticals; the last category to some extent follows Frege's idea of analyticity, although he did not restrict logic to **FOL**). We have the following strong inclusions: $\mathbf{AN}^{\mathbf{SL}} \subset \mathbf{AN}^{\mathbf{FOL}} \subset \mathbf{AN}^{\mathbf{FOL}\cup\mathbf{DEF}}$. Important differences occur between these kinds of analytic sentences. Since **SL** is decidable, we have the syntactic criterion for **SL**-analyticals (it can be regarded as an explication of Leibniz's claim that truths of reason are subjected to the method of resolution). Since **FOL** lacks decidability the syntactic criterion is not universally applicable to first-order logic. On the other hand, **SL** and **FOL** are semantically complete (A is a logical tautology if and only if A is provable inside logic). Hence, logical analyticals can be defined as universally true (true in all models, that is, valid). Furthermore, since logic does not distinguish any extralogical content, we can say that logical theorems do not provide any knowledge about specific matters (it corresponds with the point 4 of the second Nelson's proof). **FOL**∪**DEF**-analyticals do not possess such nice attributes as purely logical analytic sentences do. Since the former heavily depend on adopted definitions, they are supported by various pragmatic assumptions. Consequently, their validity is restricted to models having some intended properties. If definitions are formulated in the first-order language, corresponding theories are semantically complete. Roughly speaking, A is provable by logic plus definitions if and only if A is true in models limited (determined) by definitions.

The above considerations on three kinds of analyticals suggest that the concept of the synthetic sentence can be differently understood. If we say that $\mathbf{AN}^{\mathbf{FOL}}$ form a uniform class in spite of differences between **SL** and (first-order) predicate logic, we can very easily identify synthetic sentences as extralogical. However, this convention excludes sentences provable by logic and definitions from the set of analyticals. Thus, the statement that a female partner of a bachelor is not his wife must be qualified as synthetic. If one intends to have **FOL**∪**DEF**-analyticals, he or she needs to agree that pragmatic factors can generate analytic sentences. Call such sentences pragmatic analyticals. If they are admitted, the distinction of analytic and synthetic sentences, which is crucial for many epistemologists, becomes somehow vague. Defenders of pragmatic analyticals will probably accept that the statement about bachelors and their female partners should be considered as analytic, but the status of the sentence 'there are exactly n objects' appears as problematic. One can say that the cardinality of a model is a matter of definition, but considering the sentence in question as synthetic might be justified as well. Personally, I am inclined to recognize that pragmatic analyticals exist, the decision of this issue is not particularly relevant in the present context. Anyway, every account of analyticity, broader ($\mathbf{AN}^{\mathbf{FOL}} \cup \mathbf{AN}^{\mathbf{FOL}\cup\mathbf{DEF}}$) or narrower (only $\mathbf{AN}^{\mathbf{FOL}}$), supports the view that synthetic sentences cannot be inferred from sets consisting exclusively of analyticals. More formally, if B is synthetic, $B \notin Cn\mathbf{X}$, provided that for every $A \in \mathbf{X}$, $A \in \mathbf{AN}$. This assertion has a simple semantic justification. If $A \in Cn\mathbf{X}$, every

model of the set **X** is also a model of *A*. Assume that **X** contains tautologies only. This means that every model is a model of **X**. Since the sentence *B* as synthetic is not true in all models, $B \notin Cn\mathbf{X}$. If **X** has also pragmatic analyticals as its elements, they correspondingly limit the class of **X**-models. On the other hand, the content of *B* introduces further limitations. Consequently, the class of *B*-models is smaller than the class of **X**-models and $B \notin Cn\mathbf{X}$.

The unprovability of syntheticals from analyticals entails that no synthetic content can be inferred from the analytic content. If we agree that only synthetic sentences provide a new content (information), we justify the point 4 in (β). However, this is not the end of the story with the content of analyticals and syntheticals. Even if we consider logic as the simplest and most stable repertoire of analytic sentences, the proposed account of $\mathbf{AN}^{\mathbf{FOL}}$ requires an appeal to metalogic. One could regard the sentence (*) '$A \in \mathbf{AN}^{\mathbf{FOL}}$ if and only if $\vdash^{\mathbf{FOL}} A$' as analytic in **FOL**-metatheory, but this position begs the question because (*) does not belong to pure logic and, thereby, represents a piece of information exceeding the resources available in **FOL**. Clearly, the definition of $\mathbf{AN}^{\mathbf{FOL}}$ and its consequences are pragmatic analyticals, similarly as the statement that bachelors are unmarried men and, eventually, the sentence 'there are exactly *n* objects'. We have more complicated cases too. Undecidable arithmetical sentences, for instance, 'arithmetic of natural numbers is consistent', disregarding the view that they are sometimes regarded as synthetic a priori (see DeLong 1970, p. 222 for a discussion), can be considered as pragmatic analyticals. Some of them are true in standard models (this is the case of the assertion about the consistency of arithmetic), but other – in non-standard models (this is the case the sentence 'arithmetic is inconsistent'; clearly, the meaning of this sentence must be peculiar). In fact, the distinction between standard and non-standard models has implicit pragmatic factors. We could, for example, define bachelors as married men, but it would be extravagant according to the ordinary, that is, standard, meaning of words. The size of the metatheory used in particular clarification of analytic sentences from very rich (for instance, the method of arithmetization) as in the case of metamathematics (including metalogic) to fairly moderate as in the case of bachelors and their female partners, where legal definitions suffice. Yet some metatheoretical frameworks cannot be avoided, when we offer a clarification of the concept of analyticity or other epistemological notions. In particular, if we define analyticals as sentences which do not provide new information or even as informatively empty, our definition employs a quite definite content coming from an assumed metatheory. More specifically, the metatheory related to the concept of analyticity provides an amount of knowledge, for instance, metatheorems about **FOL** or legal definitions. Clearly, even if *A* is tautology of **FOL**, it is analytic in logic, the sentences '*A* is a tautology' and '*A* is analytic' as belonging to metalogic are not reducible to logical truths and both are pragmatic analyticals based on a quite sophisticated knowledge. Note that metatheory can also contain synthetic sentences, for instance, from cognitive psychology.

It is unclear how (β) is related to the distinction of *knowing that* and *knowing how* and conditioning the latter by the former. An answer can be suggested by (β1). Yet the assertion 'knowledge is something problematic' admits two readings

for epistemology. Firstly, one could claim that we do not know how to define this concept. Thus, the problematic character is ascribed to the notion of knowledge. Since the issue is certainly epistemological, it belongs to epistemology. On the other hand, this reading does not imply that we should abstain from accepting something as knowledge. Secondly, although we know how to define the notion of concept of knowledge (or we do not know how to do that), we have no criterion deciding what should be qualified as knowledge and what should not. Hence, the problematic nature refers to the items qualified as pieces of knowledge. The problem of knowledge becomes the issue of the validity of knowledge. It must be solved inside epistemology, not outside it. This reading explicitly ascribes to epistemology the already mentioned role of the tribunal of knowledge and, philosophically speaking, opens room for transcendental philosophy. However, it is unclear how to derive the claim that we should abstain from accepting something as knowledge from the premise that knowledge is problematic under the second reading. If we add that in order to know how to recognize something as knowledge, we must know that the piece in question is an instance of knowledge. Arguably for a transcendental epistemologist, nothing can be qualified as knowledge until epistemology fulfills its fundamental task as the authority deciding what is knowledge and what is not. The proposed interpretation of (β1) suffices to consider Nelson's second proof as going against transcendental epistemology. In fact, if we can only accept analytic sentences and the fact that the sentence '*A* is knowledge' is synthetic, epistemology cannot derive the latter from the admissible, that is, analytic premises. (β1) is relevant, because it explains why we can exclusively rely on analyticals.

The so-called presupposition-free philosophy was another target of objections advanced in (β). It is not difficult to identify Edmund Husserl as criticized in Nelson 1911, although his name does not occur in this paper. In fact, Nelson became the first serious opponent of Husserl's philosophical enterprise (see Nelson 1908, pp. 542–553). Husserl was attacked by Nelson in his earlier work mostly for the use of evidence as the ultimate epistemological criterion of the validity of knowledge and the phenomenological method in the version presented in Husserl 1900–1901. The program of philosophy as a rigorous science (that is, free of presuppositions) appeared in Husserl 1910–1911. It is perhaps interesting that remarks on the analytic/synthetic distinction in Nelson 1908 do not constitute a separate demonstration of the impossibility of epistemology, but they function rather as a supplement to (α). Moreover, this proof as formulated in Nelson 1908 has no reference, even indirect, to philosophy without presuppositions. Nelson extended his reasoning, probably intentionally in order to have a powerful weapon against Husserl (it is well known that personal relations between both philosophers were very far from being friendly; Husserl acted against Nelson's professorship in Göttingen). Nelson's critical comments asserting that philosophy free of presuppositions must be regarded as a completely unrealistic project could have been inspired by Husserl 1910–1911, but even if he had not read this book before delivering his talk in Bologna, he knew Husserl's view very well because both belonged to the Göttingen philosophical circle. Although Nelson's name is not mentioned in Husserl 1910–1911, this manifesto can be considered as a direct reply to (α) in the version offered

in 1908. Roman Ingarden, a student of Husserl, made (see Ingarden 1921) several objections against Nelson also without mentioning his name. These circumstances suggest that Nelson was *persona non grata* in the phenomenological camp.

The proof (β) demonstrates (I think that successfully, but such assertions are always risky) that presuppositions-free philosophy is actually utopian. However, it holds assuming that epistemology makes at least some synthetic statements, although its starting point must exclusively consists of analytic sentences. On the other hand, Husserl would not have agreed with such a diagnosis, because, according to his view, results of philosophical work are synthetic a priori and confirmed by apprehension of essences of phenomena. Nelson, I guess rightly, rejected Husserl's treatment of apriority outlined in *Logische Untersuchungen* as mysterious and being at odds with empirical psychological data. Transcendental phenomenology with *epoché* as its epistemic weapon did not solve troubles observed by Nelson. Thus, one could say that, *pace* Nelson, presupposition-free philosophy is either logically impossible, because it requires to infer synthetic sentences from analytic ones or recommends very suspicious modes of cognition. Both horns of this dilemma are fatal for phenomenology. Incidentally, Herman Weyl, a Hilbert's student in mathematics, but strongly influenced by Husserl on the side of philosophy argued (see Weyl 1928, and Toader 2014 for analysis and comparisons) that formalism's victory against (mathematical) intuitionism implies a defeat of pure (thai is, presupposition-free) phenomenology. Thus, if we consider the Hilbert program as close to critical philosophy, and mathematical intuitionism as phenomenology, Weyl's opinion that Nelosn's proof of (β) belongs to the same tradition. Note, however, that Weyl did not mention Nelosn in his argumentation.

The proof (β) is also interesting independently of its role as a device directed against transcendental or presupposition-free epistemology. We can and should ask whether Nelson's arguments apply to naturalized epistemology and analytic epistemology (I omit here other kinds of the theory of knowledge). In particular, I entirely omit the former and will concentrate on the latter. Since analytic philosophy proceeds by conceptual analysis, the nature of its results becomes crucial. As a working hypothesis we can assume that most assertions of analytic epistemology are analytic. In order to discuss the problem I will focus on the points (4) and (5) in (β). I recall their formulations:

(β) Propositions expressed by analytic sentences do not provide new knowledge;
(β5) If the proposition A is an instance knowledge, the proposition asserting that the proposition that A is an instance of knowledge (for brevity, $B =$ the proposition A is an instance of knowledge) is synthetic, because it is not obtained by analysis of concepts.

I will consider these assertions in the reverse order. For argument's sake, assume that knowledge is defined as a true justified belief (I do not claim that this account has no weak points). Clearly, we can consider this assumption as a pragmatically analytic one, because it is based on a conceptual analysis based on the intended meaning of words occurring in it. Take a concrete A, for instance, the already mentioned proposition that Poland became a member of EU in 2004. It is synthetic

as a piece of historical knowledge. On the other hand, if we qualify this proposition as a true justified belief, we obtain (by the rule of particularization) an analytic epistemological assertion saying that *B* (I use the earlier introduced abbreviation, but note that the latter *A* and *B* function, dependently of the context, either as metavariables or as constants referring to concrete assertions) presents an instance of knowledge. Generally speaking, we need to distinguish (a) proposition *A*; (b) a proposition that *A* is an instance of knowledge (= *B* in (β5)). The status of *A* is, at least in our example (and similar cases), indubitably synthetic, but *B* requires a further analysis. Clearly, we cannot deflate *B* to *A* by omitting the phrase 'is an instance of knowledge'. In our example, reasons for accepting the proposition that Poland became a member of EU in 2004 come from history, not from epistemology. One can call these reasons epistemic or even epistemological, but this move will not transform history into a part of the theory of knowledge. Analysis of (b) must take *A* as a granted piece of historical knowledge. Yet the occurrence of *A* in *B* suggests that this fact has some specific features as compared with its role in the historical discourse.

How to think about the proposition asserting that *A* is instance of knowledge? First of all, the corresponding sentence, that is, 'the proposition *A* is an instance of knowledge' is formally derived from the definition of knowledge as a true justified belief plus some additional information provided by empirical (or other, e. g. mathematical =investigations). In fact, it is enough to assume that the proposition *A* is a true justified belief in order to analyze the concept of knowledge by a concrete example. Is *B* expressed by a synthetic sentence? We are not forced to qualify *B* as such, because it can be considered as expressed by a pragmatic analytic sentence. The main reason is that it is obtained more by analysis of the concept of knowledge than by empirical research. This conclusion allows us to reject (β5). As far the issue concerns (β4), even if we agree that the definition of knowledge as a true justified belief appeals to an analysis of concepts and its result does not provide any new information, we can still claim that an amount of knowledge, related to how truth, justification and belief are understood, an eventually substantial empirical information surrounds our conceptual analysis. Even if we agree that *B* is an analytical, even pragmatic, do not express new instances of knowledge and do not imply synthetic assertions, a radical separation of conceptual analysis from already acquired knowledge seems fairly impossible. These observations refuse (β4). Returning to (β1), this point overlooks that doing epistemology without various metatheoretical assumptions seems impossible. This is an additional reason for rejecting the idea of presupposition-free philosophy.

Nelson's proofs (α) and (β) do not devastate analytic epistemology (or more generally, analytic philosophy). On the contrary, their analysis exhibits some essential properties of this way of doing philosophy. Firstly, results of philosophical analysis are covered by pragmatic analytical. Secondly, there are various resources of conceptual considerations. For instance, epistemology can employ various branches of philosophical logic or results of cognitive psychology. However, even if we appeal to empirical data, their embedding into philosophical vocabulary must be done. For instance, Heisenberg's uncertainty principle says nothing about determinism and

indeterminism. If one wants to use this principle in supporting an ontological view, he or she needs to embed physical ideas into philosophical ones. Similarly, Gödel's theorems say nothing about the limits of knowledge or relations between minds and bodies, but are employed in related philosophical considerations. The above remarks well illustrate the fact that extraphilosophical background of philosophical analysis cannot be established in advance or univocally determined. Some representatives of analytic philosophy prefer logic, other appeal to the ordinary language. Since resources of philosophical analysis provide various intuitions, intended or not, its results cohere with the traditional picture of philosophy as a discipline: they are equally controversial as ever have been.

Still, one lesson can be derived from Nelson. Looking for good resources justifying philosophical intuitions reminds the regressive method. In fact, when we ask, for instance, whether knowledge is closed by the consequence operation, we appeal to epistemic logic as a formal background. In other words, we regress to a formalism hoping that it helps us solve an epistemological problem. There is, however, a difference with respect to Nelson's view. He wanted to have the regressive method as converting philosophy into a legitimate scientific enterprise, and he, more or less, tried to preserves Kantian transcendentalism. The idea of philosophical analysis as outlined above has minor ambitions. It is obsessed neither by scientism nor by transcendentalism, and the objects of analysis do not need be understood as the objects of scientific knowledge or transcendental considerations. It is quite enough for many other analytic philosophers (certainly not for all) that results of analytic work can be communicated and understood by other thinkers. As usually, philosophical criticism appears as much easier than solving problems of philosophy. Although the latter aim is tempting, perhaps the fate of philosophers consists in unsuccessful attempts of achieve philosophical solutions and successful critical enterprises. Leonard Nelson is a good example in this respect.

Appendix

An anonymous referee raised some interesting question concerning possible influences of and on Nelson as well as the relation of his view to more recent philosophical discussions. Clearly, comparing Nelson's epistemological ideas in the context of old and new cognitive science could be important, but it, sa the reviewer suggests that, definitely exceeds the scope of my paper. The referee is right that Nelson's offers considerations belong to the level of metaepistemology. It, according ot the referee, invokes the question of how Nelson's ideas are related to the philosophy of language, particularly to the theory of meaning. The referee suggests to look at Nelson's presumed philosophy of language as related to what Hintikka calls 'calculism'. I would rather say that the perspective of language as universal medium vs. language as calculus is proper here. In general, Neo-Kantians can be included in the former camp, provided that the second cam considers language as asemantic. Perhaps, but it is only a very preliminary claim, Kantian and

Neo-Kantian view that metaphysics is ineffable leads to the thesis semantics shares the same fate as well. Thus, Neo-Kantians would concur with Frege in this respect. Contrary to the referee, I do not see any evidence that Nelson's problem with his professional career were related to his, eventually, unorthodox philosophical views concerning the (in)effability of semantics. Let me add that problems with Nelson's position as the ordinarius in Göttingen were not only caused by Husserl's opposition. Nelson, due to his very critical reviews of writings of distinguished German professor, including Hermann Cohen, hand troubles with obtaining PhD and Habilitation (he finally succeeded).

The referee suggests that the idea (Fries, Nelson) of justification of mathematical axioms via abstraction requires a further elaboration. I agree but I cannot say anymore than to observe that this point is unclear in Nelson. Certainly, he influenced Hilbert in pointing out the relevance of the regressive method, but I do not think that Hilbert's metamathematical arguments for adopting axioms (completeness, consistency) were suggested by Nelson. Hilbert preferred real sentences and finitary reasoning for their full intersubjectivity, but this virtue was not stressed by Nelson. He believed that we (human being) possess a kind of intuition (in Fries's sense) which leads to qualify results of regression as correct. Let we add that recent progress in reverse mathematics considering as the realization of Hilbert's program, possibly partial, more relies on purely mathematical than epistemological criteria of success.

References

Ajdukiewicz, K.: Zagadnienia i kierunki filozofii. Czytelnik, Warszawa; Engl. tr. (by H. Skolimowski and A. Quinton), Problems and Theories of Philosophy. Cambridge University Press, Cambridge (1949)

DeLong, H.: A Profile of Mathematical Logic. Addison-Wesley, Reading (1970)

Hilbert, D.: Natur und mathematisches Erkennen. Birkhäuser, Basel (1992)

Husserl E.: Logische Unteruchungen I-II. Halle: Niemeyer; Eng. tr (by J. N. Findlay), Logical Investigations I-II. Routledge & Kegan Paul, London 1970 (1900–1901)

Husserl, E.: Philosophie als strenge Wissenschaft. Logos. **1**, pp. 289–341; Eng. tr. (by Q. Lauer), Philosophy as rigorous science. In Phenomenology and the Crisis of Philosophy, pp. 71–147. Harper & Row, New York 1965 (1910–1911)

Ingarden, R. 1921, Über die Gefahr einer Petitio Principii in der Erkenntnistheorie. Jahrbuch für Philosophie und phänomenologische Forschung 4 (pp. 545–568); repr. in R. Ingarden, Frühe Schriften zur Erkenntnitheorie (pp. 48–70). Tübingen: Niemeyer 1994

Nelson, L: Die Unmöglichkeit der Erkenntnistheorie. In: Atti del IV Congresso Internationale di Filosofia. Bologna 1911, vol. I, pp. 255–275. Formignini, Genova (1911)

Nelson, L: Über das sogennante Erkenntnisproblem (book as offprint from Abhandlungen der Fris'schen Schule II, pp. 415–818 plus Register. Vandenhoeck und Ruprecht, Göttingen; repr. in L. Nelson, Gessamelte Schriften, vol. II. Hamburg, Meiner (1908)

Peckhaus, V.: Hilbertsprogramm und kritische Philosophie. Vandenhoeck & Ruprecht, Göttingen (1990)

Schlick, M.: Allgemeine Erkenntnistheorie. Julius Springer, Wien; Eng. tr. (by A. E. Blumberg), General Theory of Knowledge. Springer, Wien 1974 (1918)

Toader, L.D.: Why did Weyl think that formalism's victory against intuitionism entails a defeat a pure phenomenology. Hist Philos Log **35**, 198–207 (2014)

Weyl, H: Diskussionsbemerkungen zu den zweiten Hilbertschen Vortrag über die Grundlagen der Mathematik. Abhandlungen aus der mathametischen Seminar der Hamburgischen Universität, 6, pp. 86–88; repr. in H. Weyl, Gesammelte Abhandlungen, Band III (pp. 147–149). Springer, Berlin 1968 (1928)

Woleński, J.: Analytic vs. synthetic and a priori vs. a posteriori. In: Niiniluoto, I., Sintonen, M., Woleński, J. (eds.) Handbook of Epistemology, pp. 791–8839. Kluwer Academic Publishers, Dordrecht (2004)

Part V
Knowledge and Sciences I: Naturalized Logic and Epistemology, Cognition and Abduction

Chapter 18
Logic Naturalized

John Woods

> *"Theories of evidence and theories of knowledge are intimately linked. And there are many competing theories of evidence. One way to approach them is by looking at the theories of knowledge which are their bedrock."*
>
> (Sahlin and Rabinowicz 1998)

Abstract When logic took the mathematical turn in the nineteenth century, the human reasoner dropped out of the picture, save (at most) as a highly idealized abstraction. Although much of present-day logic retains this indifference to the realities of human cognitive agency, there has of late been no want of effort to enrich the mathematical mechanisms of formal logic in hopes of achieving a tighter fit between theory and the reasoning-behaviour of the earth-bound human agent. There is in these arrangements a clearly discernible pattern. The greater the theory's interest in approximating to how humans actually think, the more complex the theory's formal mechanisms. On this view, realist approximation varies proportionally with mathematical enrichment.

A contrary view is suggested here. It is argued that in the degree that heavily mathematicized scientific theories do well at the empirical checkout counter, their counterparts for theories of empirically instantiable and normatively assessable human behaviour are both empirical failures and normatively dubious (indeed preposterous).

An alternative approach is suggested by a nearly 50-year old development in epistemology. It is the turning proposed in 1969 by Quine in "Epistemology naturalized". The idea that is floated here is that a like transformation of logic might hold at least some of the promise that now graces philosophical work on knowledge.

Logic naturalized is an idea, not a well-worked out theoretical development. Even so, some tentative proposals are volunteered in the hope of inducing like-minded readers to consider joining the fray.

J. Woods (✉)
The Abductive Systems Group, Department of Philosophy, University of British Columbia, Vancouver, BC, Canada
e-mail: john.woods@ubc.ca; http://www.johnwoods.ca

J. Redmond et al. (eds.), *Epistemology, Knowledge and the Impact of Interaction*,
Logic, Epistemology, and the Unity of Science 38, DOI 10.1007/978-3-319-26506-3_18

Keywords Attack-and-defend networks • Causal response models • Command and control models • Consequence-drawing • Consequence-having • Data-bending • Empirical sensitivity • Epistemic dynamic logic • Fallacies • Heavy-equipment technologies • Inference-friendliness • Premiss-conclusion reasoning • Mathematicization • Naturalization • Normativity

18.1 The Mathematical Turn in Logic

It is no secret that classical logic and its mainstream variants aren't much good for human inference as it actually plays out in the conditions of real life – in life on the ground, so to speak. It isn't surprising. Human reasoning is not what the modern orthodox logics were meant for. The logics of Frege and Whitehead & Russell were purpose-built for the pacification of philosophical perturbation in the foundations of mathematics, notably but not limited to the troubles occasioned by the paradox of sets in their application to transfinite arithmetic. Logic from Aristotle to then had been differently conceived of, and would be decked out to serve different ends. The Western founder of systematic logic wanted his account of syllogisms to be the theoretical core of a general theory of everyday face-to-face argument in the courts and councils of Athens, and more broadly in the *agora* and the kitchen table. Aristotle understood that in contexts such as these premiss-conclusion reasoning[1] is an essential component of competent case-making. He thinks that when a conclusion is correctly derived from a set of premisses there exists between it and them a truth-preserving relation of *consequence*. This is a distinctively Greek idea, and one that has resonated from then to now. It is the idea that even a good deal of *everyday* case-making argument is deductively structured when good.[2] Deductivism is with us still, albeit often in rather watered-down ways. Even so, the nonmonotonic consequence relations of the twentieth and twenty-first centuries, virtually all of

[1] Readers might wonder about the concentration on premiss-conclusion reasoning. Doesn't logic also investigate higher order cognitive practices, such as decision-making, belief-change and game theoretic enterprises of all kinds? There are two reasons for the concentration. One is that the investigation of the consequences relations that underly premiss-conclusion reasoning is what logicians are best at. The other is that, in any event, most of higher level inference involves in some way or other premiss-conclusion reasoning.

[2] Although Aristotle is a deductivist about *syllogistic* reasoning, it is manifestly not his view – or a Greek one either – that all good inference is deductive. See here M.F. Burnyeat, "The origins of non-deductive inference", in Jonathan Barnes, J. Brunschevig and M.F. Burnyeat, editor, *Science and Speculation: Studies in Hellenistic Theory and Practice,* pages 193–238, Cambridge: Cambridge University Press, 1982. Reprinted in M.F. Burnyeat, *Explorations in Ancient and Modern Philosophy,* volume 1, pages 112–151, New York: Cambridge University Press, 2012. Aristotle had particular reasons for the deductivism of his syllogistic. This will become clearer later in this section.

them, are variations of variations of classical consequence. They are, so to say, classical consequence twice-removed.[3]

Whatever we might make of this lingering fondness for deductivism, the logic of premiss-conclusion inference assigns to the theorist a pair of important tasks. One is to specify the conditions under which premisses *have* the consequences they do. Call this the logician's "consequence-having" task. The second is related but different. It requires the logician to describe the conditions under which a consequence of a premiss-set is also a consequence that a human reasoner should actually *draw*. Call this the "consequence-drawing" task. The dichotomy between having and drawing is deep and significant. Consequence-having occurs in logical space. Consequence-drawing occurs in the inferer's head, that is, in psychological space.[4]

This gives us an efficient way of capturing a distinctive feature of modern mainstream logics. They readily take on the consequence-having task, but they respond ambivalently to its consequence-drawing counterpart. This ambivalence plays out in two main ways. I'll call these "rejectionism" and "idealization" respectively. In the first, the consequence-drawing task is refused outright as an unsuitable encumbrance for logic.[5] Such gaps as there may be between consequence-having and consequence-drawing are refused a hearing in rejectionist logics. However, according to the second, the consequence-having problem not only receives a hearing in logic but derives from it a positive solution. Logic would rule that any solution of the consequence-having problem would *eo ipso* be a solution to the consequence-drawing problem. The desired correspondence would be brought about by *fiat*, by the stipulation that the "ideally rational" consequence-drawer will find that his rules of inference are wholly provided for by the truth conditions on consequence itself. By a further stipulation, the conditions on inference-making would be declared to be *normatively binding* on human inference-making on the

[3]These variations on variations are more robust and complex than straightforward set theoretic restrictions. See here David Makinson's *Bridging From Classical to Nonmonotonic Logic*, volume 5 of Topics in Computing, London: College Publications, 2005. See also John Woods, *Errors of Reasoning: Naturalizing the Logic of Inference*, volume 45 of Studies in Logic, London: College Publications, 2013; chapters 7 and 8. An exception is autoepistemic consequence. Autoepistemic reasoning is discussed in chapter 10 of *Errors*.

[4]And, when vocalized, in public space or eminent domain.

[5]See Peirce; "My proposition is that logic, in the strict sense of the term, has nothing to do with how you think ....", p. 143 of Charles S. Peirce, *Reasoning and the Logic of Things: The Cambridge Conference Lectures of 1898,* Kenneth Laine Ketner, editor, Cambridge, MA: Harvard University Press, 1992, and Gilbert Harman, "Induction", in Marshall Swain, editor, *Induction, Acceptance and Rational Belief,* Dordrecht: Reidel, 1970, *Change in View*, Cambridge, MA: MIT Press, 1986; chapter 1. See also in this same vein Jaakko Hintikka's distinction between "definatory" (roughly, consequence-having) rules and "strategic" (roughly, consequence-drawing) rules in his "The role of logic in argumentation, *The Monist*, 72 (1989), 3–24, and Hintikka, Ilpo Halonen and Arto Mutanen, "Interrogative logic as a general theory of reasoning", in Dov Gabbay, Ralph Johnson, Hans Jürgen Ohlbach and John Woods, editors, *Handbook of the Logic of Argument and Inference: The Turn Towards the Practical,* volume one of Studies in Logic and Practical Reasoning, pages 295–337, Amsterdam: North-Holland, 2002.

ground.[6] There is a sense, then, in which an idealized logic closes the gap between having and drawing. Even so, it is clear that on the idealization model the gap that actually does close is not the gap between consequence-having and consequence-drawing on the ground, but rather is the gap between having and idealized drawing. In that regard, the idealization model is in its own right a kind of quasi-rejectionism. All it says about on-the-ground consequence-drawing is that the rules of idealized drawing are normative for it, notwithstanding its routine non-compliance with them. Beyond that, inference on the ground falls outside logic's remit. It lacks a lawful domicile in the province of logic.

Aristotle is differently positioned. On a fair reading, what he seeks is a new purpose-built relation of deductive consequence-having – *syllogistic* consequence – whose satisfaction conditions would coincide with the rules of consequence-drawing not under idealized conditions but rather those actually in play when human beings reason about things. Accordingly, Aristotle's is a genuinely gap-closing logic, but without the artifice of idealization. The nub of it all is that Aristotle's constraints bite so deeply that for any arbitrarily set of premisses the likelihood that there would be *any* syllogistic consequences is virtually nil; and yet when premisses do have syllogistic consequences, they are at most *two.*[7]

This is a considerable insight. Implicitly or otherwise, Aristotle sees that the way to close the gap between having and on-the-ground drawing is by reconstructing the relation of consequence-having, that is, by making consequence-having *itself* more inference-friendly. It is quite striking to modern eyes as to how Aristotle brought this about. He did it by taking a generic notion of consequence (he called it "necessitation") and imposing additional conditions on it that would effect the desired transformation. This would produce – in my words, not his – the new relation of syllogistic consequence, a proper subrelation of necessitation, whose defining conditions would make it nonmonotonic and paraconsistent, and at least some adumbration of relevant and intuitionist in the modern senses of those terms.[8] It is well to note that these inference-friendly improvements derive entirely from readjustments to consequence-having, and they put to no definitional work any considerations definitive of face-to-face argumental engagement. In other words, although inference happens in the head, Aristotle's provisions for inference-*friendliness* take hold in logical space.

[6] In a variation, normatively binding idealizations are achieved not by the logician's mere sayso, but in some putatively *à priori* sort of way, such as a conceptual analysis of the very idea of "rational agent".

[7] John Woods *Aristotle's Earlier Logic*, 2nd revised edition, London: College Publications, to appear in 2015. My thanks to David Hitchcock for helpful instruction on this last clause. I come back to this point later in this section.

[8] *Aristotle's Earlier Logic*, chapters 3 and 5. See also John Woods and Andrew Irvine, "Aristotle's early logic" in Dov M. Gabay and JohnWoods, editors, *Greek, Indian and Arabic Logic* pages 27–99, volume 1 of Gabbay and Woods, editors, *Handbook of the History of Logic,* Amsterdam: North-Holland, 2004.

When we turn from Aristotle's to modern-day efforts to improve logic's inference-friendliness, we continue to see similarities and differences. As with syllogistic consequence, the newer consequence relations trend strongly to the nonmonotonic, and many of them are in one way or another relevant and paraconsistent as well. Others still are overtly intuitionist. The differences are even more notable. I have already said that, unlike his theory of face-to-face argument-making, Aristotle's syllogistic consequence is wholly provided for without the *definitional*[9] employment of considerations about inference-making or of the beings who bring them off. In the logic of syllogisms there is no role for agents, information flow, actions, times or resources. In contrast, modern attempts at inference-friendliness give all these parameters an official seat at the definitional table of consequence-having. Consequence-having is now defined for consequence relations expressly connected to agents, information flow, actions, times and resources. There is yet a further difference to respect. It is that although these modern logics give official admittance to agents, actions and the rest, they are admitted as idealizations, rather than as they are on the ground.

In our own day, a case in point is Hintikka's agent-centred logics of belief and knowledge, in ground-breaking work of 1962.[10] Hintikka's epistemic logic is an agent-centred adaptation of Lewis' modal system S4, in which the box-operator for necessity is replaced with the epistemic operator for knowledge, relativized to agents *a*. The distinguishing axiom of S4 is $\ulcorner \Box A \rightarrow \Box\Box A \urcorner$. Its epistemic counterpart is $\ulcorner K_a\ A \rightarrow K_a\ K_a\ A \urcorner$, where "$K_a$" is read as "It is known by agent *a* that ...". We have it straightaway that the epistemicized S4 endorses the KK-hypothesis, according to which it is strictly impossible to know something without realizing you do. Of course, this is miles away, and more, from the epistemic situation of real-life human agents; so we are left to conclude that *Hintikka's* agents are idealizations of *us*. It is a gap-closing arrangement only in the sense that the behaviour of Hintikka's people is advanced as normatively binding on us. (I will have something further to say of the problem of idealized norms later in this section.)

Hintikka has an interesting idea about how to mitigate this alienation, and to make his logic more groundedly inference-friendly after all. Like Aristotle, Hintikka decides to make gap-closing adjustments to orthodox consequence-having, not just by way of specific constraints on it, but also by way of provisions that make definitional use of what agents say. That is, Hintikka decides to fit his consequence relation for greater inference-friendliness not just by imposition of additional semantic constraints, but also by application of *pragmatic* ones as well; that is, those that include a speaker's utterance conditions.

It is a radical departure. It effects the pragmaticization of consequence-having. I regard this as a turning point for most of the agent-based logics ever since. Logics of nonmonotonic, defeasible, autoepistemic and default reasoning also pragmatize

[9] As opposed to "motivational", which is another story entirely.

[10] Jaakko Hintikka, *Knowledge and Belief: An Introduction to the Logic of Two Notions*, Ithaca, NY: Cornell University Press, 1962.

their consequence relations.[11] Still, radical or not, it shouldn't be all that surprising a departure. How could it be? What would be the point of inviting even idealized agents into one's logic if there were nothing for them to do there? Consider, for example, the Hintikkian treatment of logical truth. In the orthodox approaches a wff A is a truth of logic if and only if there is no model of any interpretation in which it fails to hold. In Hintikka's pragmaticized logic, A is a truth of logic if and only if either it holds in every model of every interpretation or its negation would be a self-defeating statement for any agent to *utter*. Similarly, B is a consequence of A just when an agent's joint affirmation of A and denial of B would be another self-defeating thing for him to *say*. The same provisions extend to Hintikkian belief logics. Not only do people (in the model) believe all logical truths, but they close their beliefs under consequence. There are no stronger idealizations than these in any of the agent-free orthodox predecessor-logics.

Closure under consequence is especially problematic. In the usual run of mainline approaches, there exist for any given A transfinitely many consequences of it. Think here of the chain $A \models B$, $A \models B \vee C$, $A \models B \vee C \vee D$, and so on – summing to aleph null many in all. Take any population of living and breathing humans. Let Sarah be the person who has inferred from some reasonably manageable premiss-set the largest number of its consequences, and let Harry be the person to have inferred from those same premisses the fewest; let's say exactly one. Then although Sarah considerably outdraws Harry, she is no closer to the number of consequences-had than Harry is. They both fall short of the ideal inferrer's count equally badly. Neither of them approaches or approximates to that ideal in any finite degree. Now that's what I'd call a gap, a breach that is transfinitely wide. It is also an instructive gap. It tells us that giving (the formal representations of) agents, actions, etc. some load-bearing work to do under a pragmaticized relation of consequence-having is far from sufficient to close the gap between behaviour in the logic and behaviour on the ground.

Still, it is important to see what Hintikka had it in mind to do. The core idea was that, starting with some basic but gap-producing logic, the way to close it or anyhow narrow it to real advantage, is to do what Aristotle himself did to the everyday notion of necessitation. You would restructure your own base notion of consequence by subjecting it to additional requirements. In each case, gap-closure would be sought by making the base notion of consequence a more complex relation, as complex as may be needed for the objectives at hand. In other words,

The turn towards complexification: The complexification of consequence is the route of choice towards gap-closure and inference-friendliness.

[11] A not untypical example is any system that requires the reasoning agent to impose – if only provisionally – the closed world assumption.

One can see in retrospect that Hintikka's complexifications were too slight.[12] There is, even so, an important methodological difference between Aristotle's complexification and those of the present day. Aristotle's constraints are worked out in everyday language. Syllogistic consequence would just be ordinary necessitation, except that premisses would be (1) non-redundant, (2) more than only one and (probably) no more than two, (3) none repeated as conclusion or immediately equivalent to any other that does, (4) internally and jointly consistent, and (5) supportive of single conclusions only. These and others that derive from them would provide that the conclusion of any syllogism is either one that should obviously be drawn or is subject to brief, reliable step-by-step measures to make its drawability obvious. This is got by way of the "perfectability" proof of the *Prior Analytics*. (Even it is set out in everyday Greek supplemented by some modest stipulation of technical meanings for ordinary words).[13]

Modern gap-closers have quite different procedural sensibilities. They are the heirs of Frege and Russell, who could hardly in turn could be called heirs of Aristotle. Frege and Russell were renegades. They sought a wholesale restructuring of logic, of what it would be for, and how it would be done. Those objectives and their attendant procedural sensibilities are mother's milk for modern logicians. Logic pursues its objectives by way of mathematically expressible formal representations, subject in turn to the expositional and case-making discipline characteristic of theoretical mathematics. There flows from this a novel understanding of complexification. In the modern way, complexifications are best achieved by beefing up the mathematical formalizations of a base mathematical logic. Let's give this a name. Let's say that today's preferred route to gap-closure is the building of more mathematically complex technical machinery. In briefer words, inference-friendly logics are *heavy-equipment logics*.

Johan van Benthem has recently written of an idea that gripped him in the late 1980s:

[12]Hintikka is not indifferent to this difficulty, and seeks out some relief from it by modifying the interpretation of the K-operator. He allows it as a representation of knowledge both express and *tacit*. There is something good about this and also something less so. Cognition on the ground routinely operates tacitly and inarticulately. So it is good that an agent-based logic would take some notice of this. Not so good is the theoretical cost of the measure. Hintikka now owes us what the 1962 monograph doesn't begin to deliver; and that's a cost. Also questionable is whether Hintikka's tacit knowledge actually does effect gap-closure. Take the second incompleteness theorem as an example. We all have a Hintikkian tacit knowledge of it just because, in principle, we can all be got to see the self-defeating character of its denial without having to draw on any new empirical information. The trouble lies with the "in principle"-clause. "In principle" is justly infamous as a gap-widener, not closer.

[13]See *Aristotle's Earlier Logic*, Appendix on Categorical Syllogisms.

> The idea had many sources, but what it amounted to was this: make actions of language use and inference first-class citizens of logical theory, instead of studying just their products and data, such as sentences or proofs. My programme then became to explore the systematic repercussions of this 'dynamic turn'.[14]

In the ensuing thirty years, van Benthem and his colleagues have constructed a complex technology for the execution of this dynamic turn. It is an impressive instrument, an artful synthesis of many moving parts. Here is a close paraphrase of its principal author's summary remarks: With the aid of categorical grammars and relational algebra we can develop a conception of natural language as a kind of cognitive programming language for transforming information. This could be linked in turn to modal logic and the dynamic logic of programs, prompting insights into process invariances and definability, dynamic inference and computational complexity logics. In further variations, logical dynamics would become a general theory of agents that produce, transform and convey information in contexts both social and solo. The result is a *dynamic epistemic logic* (DEL), which gives a unified theoretical framework for knowledge-update, inference, questions, belief revision, preference change and "complex social scenarios over time, such as games." The creator of DEL also

> would see argumentation with different players as a key notion of logic, with proof just a single-agent projection. This stance is a radical break with current habits, and I hope that it will gradually grow on the reader, the way it did on me. (p. ix)

Indeed,

> . . . I would be happy if the viewpoints and techniques offered here would change received ideas about the scope of logic, and in particular, *revitalize its interface with philosophy*." (p. x; emphasis added)

Van Benthem notes with approval the suggestion in Gabbay's and my 2002 paper, "Formal approaches to practical reasoning: A survey,"[15] that the interface with argument may be the last frontier where modern logic finds its proper generality and impact on human reasoning. Again I paraphrase: Over the last decade this insight has developed into a paradigm of attack-and-defend-networks (ADNs) – from unconscious neural nets, to variations that adapt to several kinds of conscious reasoning. This, too, is a highly complex technology, a fusion of several moving parts. As provided for by Barringer, Gabbay and me,[16] the ADN paradigm unifies across

[14] Johan van Benthem, *Logic and Dynamics of Information and Interaction*, New York: Cambridge University Press, 2011; p. ix. With permission, I have drawn this paragraph and the one that follows from my "Advice on the logic of argument", *Revista de Humanidades de Valparaíso*, 1 (2013), online version at http://www.revistafilosofiauv.cl/ See also Ronald Fagin, Joseph Y. Halpern, Yoram Moses and Moshe Y. Vardi, *Reasoning About Knowledge*, Cambridge, MA: MIT Press, 1995, and Joseph Halpern and Leandro Rêgo, "Reasoning about knowledge of unawareness", *Tenth International Conference on Principles of Knowledge Representation and Reasoning,* 2006.

[15] In Gabbay et al., editors, *Handbook of the Logic of Argument and Inference*, 449–481.

[16] Dov M. Gabbay, "Equational approach to argument networks", *Argument and Computation*, 3 (2012), 87–142; Howard Barringer, Dov M. Gabbay and John Woods, "Temporal argumentation

several fields, from logic programs to dynamical systems. AD-networks have some interesting technical capacities. They give an equational algebraic analysis of connection strength, where stable states can be found by way of Brouwer's fixed-point result. When network activity is made responsive to time, logic re-enters the picture, including the development of quite novel modal and temporal languages. "Clearly", says van Benthem, "this is an immense intellectual space to consider." He adds that he "totally agrees" with the ADN "vision, and am happy to support it." (p. 84)

Here, then, are just two of a great many heavy-equipment technologies, specifically adapted to the requirements of argument. They are unifications of partner elements, some of their authors' own contrivance, but in the main having an already established and well understood methodological presence in the several research communities from which they have been borrowed. Both the DEL and ADN approaches carry the same presupposition for the logic of argument, and underlying it the logic of inference too. It is that argument and inference won't yield the mysteries of their deep structures unless excavated by heavy-equipment regimes capable of mathematically precise formulation and implementation. It is here that the fissure between modern logic and Aristotle's is deepest and most intensely felt, at least by me.[17]

Why, then, it might well be asked, my *own* complicity in constructing ADN logics and the formal models on display in earlier work?[18] I am not opposed to heavy equipment methodologies as such. I am perfectly happy to see formally contrived new ideas added to our conceptual inventories, for whatever good may be in them in due course, apart from their beauty as intellectual artifacts. I also concede the necessity of idealizations, formally wrought or not – even those that are transfinitely untrue to what happens on the ground – that are indispensable to the *descriptive* success of the empirical sciences; not excluding those of them that investigate empirically realized and normatively assessable human goings-on *in terra firma*. I also welcome the fact that thinking of things in ways they couldn't possibly be sometimes gets us to thinking of things, even perhaps of other things, in ways they

networks", *Argument and Computation,* 2–3 (2012), 143–202; and Barringer, Gabbay and Woods, "Modal argumentation networks", *Argument and Computation*, 2–3 (2012), 203–227. See also Artur d'Avila Garcez, Dov M. Gabbay, Olivier Ray and John Woods, "Abductive reasoning in neural-symbolic systems", *Topoi*, 26 (2007), 37–49, and Artur d'Avila Garcez, Howard Barringer, Dov M. Gabbay and John Woods, *Neuro-fuzzy Argumentation Networks*, to appear.

[17] John Burgess reports the standard view that "formal logic" is a pleonasm and "informal logic" an oxymoron. See his *Philosophical Logic*, Princeton: Princeton University Press, 2009, p. 2. In the late 1970s, informal logicians used to fret about the suitability of the adjective "informal" in apposition to so noble a noun as "logic". "Informal", they feared, bespoke a kind of casualness or, heaven forbid, sloppiness. One day Michael Scriven stated his own view of the matter. "Informal logic" was indeed the wrong name for their enterprise. "But what should we call it?", he was asked. "Call it what it is", said Scriven. "Call it '*logic*'".

[18] See, for example, Dov M. Gabbay and John Woods, *Agenda Relevance: A Study in Formal Pragmatics* and *The Reach of Abduction: Insight and Trial*, volumes 1 and 2 of *A Practical Logic of Cognitive Systems*, Amsterdam: North-Holland, 2003 and 2005.

do turn out to be.[19] In this present section, I've been trying to set my course for the developments that lie ahead in section II. Part of what I want to say is how much I distrust our present compulsion to mathematicize everything in sight. Compulsions aren't good for intellectual health. They are a drag on the market and a pathological impediment to open-minded enquiry.

A further reservation concerns the groundlessness of the pretensions of the heavy-equipment methodologies to a *normative authority* over human cognitive performance in London, Vancouver and Guangzhou. The two most noticeable explanations of the normative authority of ideal models are the *reflective equilibrium* defence, and what I'll call the *mathematico-analytic* defence. According to the first, the correct procedures for action are those implicitly in play in the relevant community of agents. The trouble with this is the impossibility of finding credible candidates to qualify as relevant communities. If it is the human community on the ground – that is, all of us in general – then there is between how we perform and what the orthodox models require us to perform no equilibrium at all. On the other hand, if the authoritative community is the ideal-modelling research community, there will indeed be a nice concurrence between what their models demand and what they say should be demanded. Which prompts a good, if somewhat informal, question: "Who made these guys king of the normativity castle?" Besides, why would we think that *saying* is a salient consideration? What the experts say at the office is one thing. In all other respects, they are just like the rest of us.

The mathematico-analytic defence is even more of a muddle. In one version of it, an idealized norm is behaviourably binding on the ground when it arises in the theory as a theorem. In another, the norm's authority arises from the fact that it is analytic in the model (i.e. made true there by stipulation or nominal definition). The general idea is this: The proposition "$2 + 3 = 5$" is a theorem of number theory; and some people think that it is true by meanings alone. Its normative authority is straightforwardly clear. If someone in London, Vancouver or Guangzhou wants to add 2 and 3 in the way authorized by number theory, he should not identify their sum as any number that's not 5. The same would be true for belief-closure. If someone on the ground wanted to close a belief in the way authorized by idealized closing, he would fire away transfinitely. Of course, this is absurd. No one on earth, except for the odd decision-theorist when at the office, has ever heard of the idealized closure-conditions, never mind aspiring to their fulfillment.[20]

From the very beginnings and most of the time thereafter, the logician had to be two things at once. He would be the setter of the targets for his theory, and he would be the creator of the tools to enable him to meet them. If we were speaking of cars

[19]The idealization that populations are infinitely large plays an indispensable role in population genetics. No one thinks that population genetics tells us anything at all about the cardinality of empirically realizable populations. But everyone knows that it tells us a good deal about natural selection on the ground.

[20]For a more detailed consideration of these points readers could consult my "Epistemology mathematicized", *Informal Logic,* 33 (2013), 292–331, and *Errors of Reasoning*, chapter 2, section 3.

rather than inferences, we could see this duality nicely captured by a quite common division of labour in Detroit. Cars would be sold by the sales staff, but they would be built by the engineers. Things are different in the logic business. Not many of Ford's sales people know anything much of how cars are built, and engineers are notorious for their poor salesmanship. But in logic, it falls to the logician to build what he sells. He must be his own engineer. It is not at all surprising that Ford's top salesman might know nothing of engineering. But the same thing in logic would be quite astonishing.

There is a further difference between the car business and logic. Logic's modern machinery is put together in a quite particular way. Originally designed for expressly mathematical purposes, its creators, then and now, bring a generalized mathematical sensibility to their creative work. In due course it would become apparent that the technical objects of their machinery are themselves possessed of a mathematical character and are eligible for mathematical investigation in their own right. In the car business the work of the engineering division and the work of the sales division is harmonized by the biting discipline of the bottom-line. No engineer will thrive in Dearborn if the company's cars don't sell, even if he's more interested in new equipment than he is in new cars. Logic is different. By and large, the work of logicians is free of commercial expectation.[21]

When we put these two points together, we can see a quite considerable alienation of the mathematical study of logic's machinery from the question of what the equipment might be good for. The factor of good-for recedes into the background, and technological self-study becomes *sui generis,* and withal the route to the upper elechons of academic achievement and repute.[22] The heavy equipment logics of the day have put themselves in an awkward position. They *say* that their technical complexifications are wanted for the inference-friendliness. But they construct their complexifications in ways that discourage if not outright preclude the accomplishment of those ends.

With complexification comes complexity, which is a well-known inhibitor of on-the-ground implementability.[23] This necessitates the reinstatement of idealizations, for two particular reasons among others. Idealizations would simplify and

[21]This is less so in departments of computer science. It is an interesting story to tell, but longer than there is space for here.

[22]It would be rude to speak here of heavy-equipment autoeroticism, but we would know what was meant.

[23]There is a significant ambiguity between something that can be implemented *in* a human agent, for example, a model that is realized in a neural net, and something that can be implemented *by* a human agent, as for example, a rule of decision manoeuvre that is simple enough for a person to deliberate upon and follow. Complexity is much less a problem for the first than the second. Nevertheless, the purport of heavy equipment gap-closers is to facilitate implementability-*by.* For implementability-in, see again Gabbay *et al*., *Neuro-fuzzy Argumentation Networks.* On the other hand, *Errors of Reasoning* offers an alternative: For implementability-by, it is advisable to drop the idea that real-life reasoning competence is intrinsically, or even generally, a matter of following *rules*.

streamline procedures for theorem-proving; and they would explain the broadening gap between having and drawing occasioned by the idealization process itself. This would be brought about in the same old ways: by closing the gap between having and drawing in the heavy-equipment model, and by normativizing the model's drawings in relation to those that play out in the world. This is seriously problematic. Heavy equipment upgrades yield empirically false accounts of on-the-ground drawing, and do so in ways that exacerbate, not solve, the normativity problem. (As for my own involvements with the heavy-technology industry, I have never supposed that the ADN technology is normatively authoritative for anything apart from its own self-created objects. The same holds, I believe, for my ADN co-conspirators.)

One thing that could be done – and in some cases has been – to mitigate the gap-producing difficulties engendered by ideal models is to deny the transfinitely false ones a place at the table, and admit only those falsities for which an approximation relation is either definable or at least plausibly entertainable. With belief still our example, closure under consequence would not be permitted, but a sizeable gap could still remain between drawings in the model and drawings in human life. The difference would be that, where the original gap is transfinitely wide, the new gap is smaller – anyhow smaller enough to qualify the new closure-rule as approaching in some finite degree what actually happens here.

So adjusted, the heavy equipment approach to inference-friendliness could now be roughly summed up this way:

Complexity as gap-closing: The heavier the equipment the less empirically unfaithful the machinery's formal models, to the degree that they *approximate* to what happens on the ground.

It is an interesting idea, animating another.

Approximation converges on normativity: The closer the approximation of a theoretical model of premiss-conclusion, the greater its descriptive adequacy; and the greater too, its presumptions of normative sway.

As far as I can tell, nothing in the heavy equipment literature puts things in just this way, or even close to it. And a good thing, too, readers may be thinking! Isn't everyone still cringing at *le scandale* created by poor Mill's gaffe in proposing in *Utilitarianism*, chapter 4, that "the sole evidence … that anything is desirable, is that people do actually desire it"? A not uncommon complaint can be found in Charles Hamblin's observation that "[i]t was given to J.S. Mill to make the greatest of modern contributions to this Fallacy [= the 'naturalistic' fallacy] by perpetrating a serious example of it himself."[24]

The mockery is misplaced. It is little more than name-calling, occasioned by the critics' misconception that Mill is saying that "The desirable (F) is what's normally desired by us all (G)" is true by meanings, supplemented by the further assumption

[24]C.L. Hamblin, *Fallacies*, London: Methuen, 1970.

that believing that something is F entails believing that it is G. The first of these assumptions is implausible on its face. The second owes its truth (if true it be) to the falsehood that believing that this thing *a* is F requires that there be some distinct term "G" that the believer in question believes to be semantically equivalent to "F". Notwithstanding the stout resistance it provokes, the convergence of approximation on normativity is an extremely engaging idea, whatever its prior origins. It carries a suggestion of the first importance for the logic of consequence-drawing. It is that the normative authority of a logic converges on its descriptive adequacy. Should this prove to be so, it deserves acknowledgement as a foundational insight for a naturalized logic of inference.

18.2 The Naturalistic Turn in Logic

It is easy to see that the idea that a theory of premiss-conclusion reasoning's normative authority varies in some nontrivial way with its descriptive adequacy bears no *intrinsic* tie to theories contrived in the heavy-equipment way. So it would be a mistake to think that a theory's approximation to empirical fidelity is owed to its mathematical complexity. It lies instead in the fact (if fact it be) that its still-considerable falsities remain eligible for the duties of real-life approximation. It is too early to declare the heavy-equipment industries a dead loss for inference-drawing on the ground. But it surely can't hurt to start looking for alternatives.

I have an idea about where the search might pay off. It might pay off in a logic reconstructed in the way that traditional epistemology was adjusted by Quine and others in 1969 and following. Just as Quine proposed the *naturalization* of epistemology, so I now propose the naturalization of *logic*.[25] The pivotal point of it all is this. What is the use of admitting to one's logic agents, actions and the like, if we don't admit them as they actually are on the ground – warts and all? Of course the last thing that Quine would ever agree to for logic is to do what he himself did for epistemology. For Quine, logic was first order classical quantification theory and nothing else. Quine wanted no truck with people in his logic, formally idealized or in the flesh. He clung to this conservatism until late in his career, when he grudgingly allowed that physics might require a nonclassical quantum logic, and constructivism in mathematics an intuitionist one.[26] The idea of naturalizing logic does not originate with me. In the modern era alone, it is actively proposed by Dewey and sympathetically entertained by Toulmin and Finocchiaro.[27] In this

[25] "Epistemology naturalized", in *Ontological Relativity and Other Essays,* pages 114–138, New York: Columbia University Press, 1969.

[26] I am just one of the many who think Quine's contributions have done epistemology nothing but good. This makes me hopeful that a like success might be had with logic.

[27] John Dewey, *The Later Works,* 17 volumes, Carbondale: Southern Illinois Press, 1981–1991; volume 12, 27; Stephen Toulmin, *The Uses of Argument*, Cambridge: Cambridge University

second section, I'll give a sketch of how the naturalization of logic might go, with some indication of the good of it. To this end, I'll call upon the reflections of Nat, an imaginary would-be naturalizer.

Nat is a young logician raised in the heavy equipment tradition, and a great respecter of it. But he has grown sceptical of late about the normative legitimacy of complexly mathematicized models those kinds of behavior that are not only empirically instantiated but susceptible as well of *performance-assessment*. He has also become doubtful of the idea that models that formally represent human behavior serve, just so, as well-regulated approximations of it. Nat is aware that there were times when Quine – the great naturalizer of epistemology – was drawn to the idea that the best way to proceed was to stop doing epistemology and to replace it instead with psychology. As regards the naturalization of logic, Nat has never thought that psychology should replace logic. The naturalization he seeks is not a replacement naturalization; it is a cooperation naturalization. For someone in Nat's position, this is problematic. Nat is a logician, not a psychologist, although he had worked hard to get himself up to date in the philosophy of psychology. Still, it was not clear to Nat how a productive partnership with psychology could be effected by a logician. He is no less in the dark about how it might be effected by a psychologist. However it would be done, Nat would be well-advised to proceed with caution. In time, he came to a procedural decision: He wouldn't in the first instance seek an acquaintance with psychology's leading *theories.* He would concentrate at first on the *data* that those theories collect and classify.

Nat thought he could hold a naturalized logic to an adequacy condition of "empirical sensitivity". To bring it off, the logician would familiarize himself with the data that the cognitive sciences seek to account for. He would in time develop an informed acquaintance with the findings of the empirically best-confirmed of those theories. The naturalizer would offer an account of any of his own logic's empirical disconformities with the data and findings of the partner sciences. Nat was also quick to appreciate a fact openly on view on the ground if only we would take the trouble to look. It is that the human animal is a knowledge-seeking organism and that premiss-conclusion reasoning is an important facilitator of its achievement. Accordingly, in addition to its empirical sensitivity, a naturalized logic of reasoning should display an *epistemic* sensitivity as well. This is already shaping up to be a coherent methodology, beginning with a respect-for-data principle.

Respect for data: Logic should develop a healthy respect for the data and should respond to them with empirical and epistemic sensitivity. And correlatively it should respect the *difficulty* of paying them proper respect.

No doubt some readers will think this a platitude, and a condescending one at that. This is half-right. It *is* a platitude to emphasize the necessity to respect the

Press, 1958, 257; and Maurice Finocchiaro, *Arguments About Arguments*, New York: Cambridge University Press, 2005, 6–7. I first became aware of Finocchiaro's naturalistic leanings when viewing a poster for his 1987 lecture at the University of Groningen, entitled "Empirical logic".

data. What makes it a platitude worth voicing is the second thing it asserts. Getting the data right is as a matter of course more difficult than we might think. So no condescension is intended.

When Nat speaks of a theory's data he has one or other of three things in mind. Sometimes they are the data that fix the theory's subject matter. Sometimes they are the data summoned up for the purpose of confirmation or disconfirmation. At other times, data are understood as pretheoretical beliefs about the prior two – prior beliefs about the enquiry's subject matter and beliefs about the theory's test data, sometimes referred to as "our preanalytic intuitions" about the matters to which the theory will turn its attention.

Nat is trying to imagine a way of deciphering and organizing data as freely as possible from preconception, which is precisely what pretheoretical belief happens to be. In this he is not much influenced by the fact that data always underdetermine theories, for that is little more than the fact that the data that provide a theory's confirmation do not, in so doing, logically imply it. Nat's is a more practical worry. Here is a case in point. Nat sees the naturalized logician as studying premiss-conclusion reasoning as it plays out in the various scenarios of human life. To do this, he will need to collect and classify some subject-matter data, including some premiss-conclusion reasoning behaviours. Where will he find this behaviour? One possibility is that it will be found in the *linguistic* behaviour of the human agent, in utterances containing trigger-words such as "therefore", "since", "it follows that", and the like. Linguistic behaviour presents itself in one of two ways. It can occur spontaneously, that is, independently of the investigator's interest in it. Or it can be elicited, stimulated by the investigator's questions about what the subject is able to tell him of his own premiss-conclusion reasoning experience, and perhaps, too, of others as well.

Nat has a number of reservations about these suggestions. One is that, judging from his own case, most of a human being's premiss-conclusion reasoning occurs unvoiced. Another worry is that much of the time when we give voice to our premiss-conclusion reasonings, our motivation is not reportorial but dialectical. We aren't giving voice to our reasonings; we are *defending* them. A third hesitation concerns what we are able to report when prodded by an investigator. Suppose he asks, "What is it about p that makes you think that, if true, q might well be true too, but not r?" For Nat, the plain facts are these. It is hardly ever the case that this is a question we are able to answer, except at most clumsily. Hence any answer preferred is likely to be irrelevant or just wrong.

Nat has a further hesitation. About elicited linguistic behaviour. Suppose the investigator asks the subject group whether a given piece of premiss-conclusion reasoning is correct – for example, "It's an all-day rain, so they won't go ahead with the picnic." Suppose the answer is No and that the reason for it is that it is not valid. That, thinks Nat, would be a case in which the subject's pretheoretical belief about reasoning-adequacy embodies a faulty preconception.

Consider in this regard Gerd Gigerenzer's alarm about the "data-bending" he sees in the application to cognitive processes of methods developed by theories

of statistical computation and theory testing in the 1940s and 1950s. Under the influence of such assumptions, theories of cognition in experimental psychology

> were cleansed of terms such as restructuring and insight, and the new mind has come to be portrayed as drawing random samples from nervous fibers, computing probabilities, calculating analysis of variance, setting decision criteria, and performing utility analyses.[28]

Taken together, these assumptions conceptualize the human reasoner as an "intuitive statistician". This "radically changed the kind of phenomena reported, the kind of explanation looked for, *and even the kind of data that were generated*."[29] What is more, researchers who adopted the methods of inferential statistics were unaware of this change, since these methods had become canonical in psychology.[30]

Gigerenzer's defection from the intuitive statistician orthodoxy is a hotly contested development in present-day psychology, with Gigerenzer's still very much the minority position. I cite Gigerenzer not to endorse him on the particulars of the intuitive statistician hypothesis, but rather to acknowledge data-bending as a general ill to which all empirical science lies exposed. Data-bending is misconstrual of how the data actually are, owing to the wrong sort of data-loading assumptions. It is conceiving of facts on the ground in ways that facilitate pre-conceived theoretical outcomes. This is nothing to make light of. We have known at least since Bacon that data aren't self-announcing, and yet that they can't be grasped at all without some prior or concurrent conceptualization of them. This generates a nasty problem for the experimental theorist. He can't proceed without conceptualizing his data, and yet the line between conceptualization and misconceptualizations is easily transgressed. There are no algorithms for the avoidance of data-bending. But there are some useful lessons to be learned. One is to be careful. Another is to respect these difficulties. See here the admonitions contained in Patrick Suppes' classic paper "Models of data".[31] Suppes is essential reading for the would-be naturalizer.[32]

Nat attaches a singular importance to these lessons. He also views with suspicion the most prominent justification for sticking with an empirically false theory. According to that view, it is not the aim of such theories to be descriptively adequate; the goal is to establish rules that are normatively authoritative for human practice on the ground. Fundamental to Nat's project is his rejection of this assumption, at least until such time as it might come to have a convincing independent defence. Like

---

[28]Gerd Gigerenzer, "From tools to theories", in Carl Graumann and Kenneth J. Gergen, editors, *Historical Dimensions of Psychology Discourse,* pages 336–359, Cambridge: Cambridge University Press, 1996; 339.

[29]Idem; emphasis added.

[30]Further details may be found in Dov Gabbay's and my "Filtration structures and the cut down problem for abduction", in Kent A. Peacock and Andrew D. Irvine, editors, *Mistakes of Reason: Essays in Honour of John Woods*, pages 398–417, Toronto: University of Toronto Press, 2005; 411–414.

[31]In Ernest Nagel, Patrick Suppes and Alfred Tarski, editors, *Logic, Mehodoloy and Philosophy of Science: Proceedings of the 1960 International Congress,* pages 252–261, Palo Alto: Stanford University Press, 1962.

[32]Further discussion can be found in *Errors of easoning,* sections 14.1–14.4.

the rest of us, Nat knows full well that science frequently adopts false idealizations, such as the infinite cardinality of populations in population genetics. These are, he says "virtuous distortions",[33] and are so when they are compensated for by the confirmation of the theory's observational predictions. Normatively idealized theories of empirically realized and normatively assessable human performance can't meet that test – consider here the perfect information assumption for ideally rational decision systems. So the theory's only payment alternative would appear to be its success at the *normative* checkout. But Nat's view is that there is little in the way of a settled consensus about what doing well there would actually amount to. In the absence of a sounder grasp of normative legitimacy, Nat thinks that the better option is to try to handle the problem of empirical infidelity in a different way.

Nat's own response is twofold. He stresses the priority of attending with care to the actual details of human reasoning on the ground, without preconception and well in advance of assessments of goodness or badness. And his worry about preconceptions is an ecumenical one. He dislikes the preconceptions of orthodox logic. But he dislikes no less the preconceptions of orthodox psychology.[34] He is distrustful of scientific disciplines that organize themselves into "schools".

Nat is starting to have a working idea of how he thinks the naturalization process should go. To this end, he has availed himself of a fable, a kind of thought-experiment about the ways of naturalization. Nat imagines that for some years now a visiting team of cognitive anthropologists from a distant and unknown extraterrestrial place has been hard at work in our midst. Its earthly mission is the examination of human cognition. These cosmonauts from afar are well-equipped for their work. Themselves organic beings, they have been able to acclimate to the particularities of planet Earth. Themselves cognitive agents, they have some acquaintance with how cognition works under the ecological constraints of habitat. They are also accomplished field linguists and intelligent problem-solvers. Nat has adapted this fiction of the visiting scientist from van Fraassen's notion of an epistemic marriage, which envisages an epistemic partnership between dolphins, extraterrestrials and us.[35] Just as Wittgenstein would wonder whether if a lion could talk, we would be able to understand him, the visitors were interested in whether

[33]John Woods and Alirio Rosales, "Virtuous distortion in model-based science", in Lorenzo Magnani and Walter Carnelli, editors, *Model-Based Reasoning in Science and Technology: Abduction, Logic and Computational Discovery,* pages 3–30, Berlin: Springer, 2010.

[34]Nat has heard and liked an amusing bit of mischief about physicists. It is said that physicists find fault on two counts with biologists. One is that biologists aren't very good at data-analysis. Another is that they aren't particularly adept at model-building. Of course, it's only a story, a kind of jape really. But it strikes Nat that, given the marked difference in the respective complexities of the motivating and confirmatory data of these two disciplines, perhaps the story has something of the ring of truth. It also strikes him that if this were really so, how less it could be so for the *social* sciences?

[35]Bas van Fraassen, "The day of the dolphins: Puzzling over epistemic partnership", in Peacock and Irvine, *Mistakes of Reason*, 111–133.

our own information-processing activities rose to the bar of cognition in a manner that would be discernible to their own methods of enquiry.[36]

The visiting anthropologists are here to do some descriptive epistemology, to take the earthly cognitive pulse as best they can with the exploratory resources available to them. They will run their investigations in compliance with their own understanding of best scientific practice. They will leave themselves free to consult the published record of *our own* conception of best scientific practice, but without prior commitment to defer to ours when it conflicts with theirs. Neither will they consult our philosophers, not even the ones who do logic; at least not until their own work is finished. Why would the visitors bother with philosophers? They are not themselves philosophers; they are scientists conducting a naturalistic examination of the naturally occurring phenomena of cognitive behaviour in their subject population. This exclusion is neither ideological nor hostile. It simply reflects the plain fact that scientists do their thing and philosophers do their largely different thing. Philosophers and scientists can share a common subject-matter, but they each tend to cut their respective cakes largely in independence of the other. A sensible course for each would be to arrive at a decision about what happens now. There might be some advantage for each in exposing one another's views of the matters they have in common to a reciprocal scrutiny. But the visitors' point, and Nat's too, is that this scrutiny is better reserved until the scientists have done their business.

The visitors' first task was the identification, classification and description of their enquiry's data. The data they sought were the observable manifestations of knowledge-seeking, knowledge-attaining and knowledge-transmitting human behaviour. For all the disadvantage that accrues to their status as aliens, they were free of any of the preconceptions and confusions specific to human enquiry. The same could not be said for the preconceptions and confusions specific to their own ways of proceeding. This left them no choice but to proceed with the greatest circumspection in the collection and analysis of data. They were respecters of the Respect for Data principle virtually by default. Still it's not true that the visitors had nothing to go on apart from their own understanding of how enquiries should be conducted and their own considerable respect for the difficulty of the data analysis task. They came equipped with working assumptions of no mean significance. One was that the physical construction of the human animal made it more than likely that they were built for and capable of knowledge. The other was that, given the highly social character of human ecologies, it was safe to assume that humans themselves were at home in the recognition of observable signs of human decision-making, belief-revision, problem-solving, knowledge-seeking, and all the rest. So the newcomers had a stake in identifying the behaviour that attended the efficacious exercise of these *human* skills.

[36]The literature on lionspeak and related matters is large, and amply cited in Dorit Bar-On's "Expressive communication and continuity scepticism", *Journal of Philosophy,* 60 (2013), 293–330.

No doubt some readers will dislike the story of Nat and the visitors on methodological grounds. They will see it as unscientific, as the displacement of honest intellectual work by make-believe and wool-gathering. It is, in fact, no such thing. My task is to try to imagine what it would take to put into effect the naturalistic programme under the constraints I've imposed on it. One of those constraints calls for data collection and data analysis as free as possible from theoretical preconception and pre-emptive conceptualization. But it is precisely these traits that the human animal has in abundance and is attached to at the hip. Nat is like this too; he is one of us. He, too, is trying to imagine how he would go about the job of respecting the data if he were unencumbered by the dispositions that encumber all his fellow beings. So he tries to imagine how intelligent *non*-humans who, lacking the preconceptions and confusions distinctive of humans, would go about the business of finding the data on which to erect their account of how the human animal sets and discharges his cognitive agendas.

So, then, Nat is not larking about. His fable has a serious purpose. Nat knows that any human being who sets out to make the observations he seeks could only do so in *medias res*, hence as an inheritor of the received wisdom, as of then, as to how those things should go, and about what is already known about them. Try as he might, Nat is having a tough time in shedding his own orthodoxies. He wants something closer to a clean-slate way of proceeding.

The visitors decided that it was not part of their mission to discern how on-the-ground human inference-makers *defined* inference – or knowledge or belief or whatever else. Nor would they circulate questionnaires asking their subjects to tell them whether they thought that this, that or the other thing is *bona fide* knowledge or good reasoning. The visitors have long been aware that, among their own kind, people do quite well at knowing and reasoning without being very good at saying what these things are. Until they learned otherwise, they would assume the same for us. From this came another foundational insight:

Being good at and knowing what: Being good at knowing things and reasoning well does not require knowers and reasoners to know how to define what knowledge and good reasoning are, or to specify the conditions that bring these things about.

Substitution here for "reasoning" of "remembering", "imagining", "seeing", "high-fly ball-catching" and a plethora of others generates a host of statements most of us would consider too obvious for words. Consider, for example, our knowledge by looking that it has snowed overnight. Everyone in town with eyes to see knows this to be so. But hardly any of them has much of an idea about how the mechanics of visual cognition actually play out. Why wouldn't this also be true for "knowing" and "reasoning"?

Perhaps it will be protested that if we are so good at knowing things, why are we so bad at knowing what knowing is? It is an excellent question but a feeble protest. Does the questioner really think that his question carries the force of rebuttal? If so, he would do well to explain why. When, in the late stages of their enquiry, the visitors relaxed their exclusion of earth-bound philosophy, they were astonished to learn – disapprovingly so – that some of the first of our great philosophers were in

thrall to the idea that a person's knowledge requires a prior or concurrent grasp of a real definition of it, that a concept can't be instantiated in human behaviour in the absence of the behaver's "analytic grasp" of it. This, the visitors thought, amounted to a scepticism so corrosive as to make the attainment of knowledge a generally impossible target. When some of our own earthbound professors of epistemology protested in turn the excessiveness of the visitors' alarm, they answered in unison: Socrates and *les autres* could not have thought this way had they paid scrupulous attention to the ups and downs of what really goes on in human life; especially in the *agora.*

Perhaps the visitors were a trifle hasty. They didn't have time to immerse themselves in the history of earthly epistemology. But they'd had enough exposure to it to have tasked themselves with the four questions. One is whether having a real definition – or conceptual analysis – of it is a condition on knowing what knowledge is. A second is whether knowing what knowledge is a condition of there being any. The third is whether the ancients were inclined to favour affirmative answers. The fourth is whether this favoritism has a discernible presence in modern-day analytic approaches to epistemology. Not having the time or inclination for extended consideration of these matters, they came together on three summary positions; *first*, that any notion that a scientific knowledge of knowledge is to be got by a conceptual analysis of "knows" is a misbegotten idea; *second* that the only room for big-box scepticism in the science of human knowledge is by way of the default rule that big-box scepticism in the science of human knowledge should not be so much as entertained, never mind rebutted, except for weighty cause; and *third* that the earthly epistemological tradition betrays too little heed of these constraints; not without exception, but dominantly so. They reckoned that a thoughtful examination of the cognitive routines of human life make it clear that just about the last thing that could be true of it is radical scepticism of any broad kind. On the contrary, they thought, it was empirically evident that human beings are good at knowing things – not perfect but good; that they have lots and lots of it about lots and lots of different things; that the human cognitive harvest is both abundant and diverse.

A further shock was administered when, shortly before heading for home, the group began to look into what earthbound *logicians* had been saying about these things. They were taken aback to discover the confidence and wide-spreadness of the dogma that premiss-conclusion reasoning is no good when it fails to be deductively valid or at least inductively strong in the technical sense familiar to the statistico-experimental sciences. What surprised them most was the utter lack of behavioural recognition of this would-be fact in the subject population. Most of the reasoning that passes muster there – and is evidently accepted as good – fails both these standards. This left the visitors with two choices. They could condemn their human subjects as across-the-board losers in the reasoning game. Or they could reject the validity-or-inductive strength condition as a general requirement for good premiss-conclusion drawing. Of course, they chose the latter. How could they have not? They were natural scientists who faithfully respected the necessity of respecting the data on the ground. I hardly need say that Nat was quick to sign on, and in short order would see that one of the essential tasks for a naturalized logic of inference

is the discovery of the conditions under which this "third way reasoning" would be properly achieved; and in the process to see that the resulting third way logic would be the natural home for virtually all the nonmonotonic logics currently on offer, once such compensating adjustments as might be required were worked out.[37]

It was equally clear to the visiting team that there is another abundance for the naturalized logician to pay attention to in the cognitive ecologies of the subject population. Knowledge exists there in abundance. Error is another of its abundances. The human animal makes lots and lots of errors about lots and lots of different things. A logic of reasoning must bring these abundances into a benign harmony, in response to yet another empirically discernible feature of human cognition.

Enough-enough: Human beings know enough about enough of the right things enough of the time for survival and prosperity and, from time to time, for the erection of civilizations of dignity and lingering worth.

There is no question here of our prosperity *entailing* an abundance of enabling knowledge. Cognitive abundance is our visitors' working hypothesis. It is not ruled out that there is none better. It is not ruled out that they might come to think that a more plausible working hypothesis would be that the beliefs that serve us well are mainly false, in which case, *right* belief would separate away from *true* belief. But, if anything was clear to the anthropologists, it was that beings like us are awash in what we might call *alethic* beliefs, which are believings-to-be-true. This gives the new working hypothesis a bit of a twist, providing that the best way of achieving prosperity is believing-to-be true propositions that are false. This handed the visitors a chuckle and the idea was dropped like a hot potato. Where, they asked, is it empirically discernible in the belief-prosperity data that this story would hold in the general case?

Nat followed the visitors in thinking that an attentive naturalizer would lodge his curiosity about good and bad reasoning in a default principle for the logical theorist:

NN-convergence: Take it as given in the absence of particular reasons to the contrary that humans reason well when they reason in the ways that humans normally reason in the conditions of real life. That is, good premiss-conclusion reasoning is reasoning as usual. The *normative* converges on the *normal.*

Nat was careful (and well-advised) to add an important caveat. The NN-convergence principle is *not* a safe default for all aspects of human cognitivity; its application here is reserved for premiss-conclusion reasoning. It tells us to judge premiss-conclusion reasoning in roughly the same sort of way that we'd check the subject's pulmonary behaviour. What this amounts to is that, for the most part, a human agent's premiss-conclusion reasoning is the right way to reason when his conclusion-drawing mechanisms are in good working order, and at present working in the right way, engaging good information in the absence of hostile externalities.

[37] Still, in so saying, Nat was not giving up his renunciation of the orthodox approaches to reasoning even where deductive validity or inductive strength is the appropriate assessment-standard.

Breathing is like that too. By and large, the goodness of good breathing depends on the ship-shapeness of the pulmonary equipment. At bottom, as we might say, good breathing is not down to us; it is down to our equipment. Sometimes, however, it *is* down to us. We can't sing grand opera, or achieve a place on the Vancouver Canucks, if we don't breathe in the required ways. We can't bring this off until we learn to do it. Almost always we'll have to be taught by an expert, and oftener than not it won't work. Hardly anyone has the breath for *Tosca* or for the seventh and deciding game of the Stanley Cup playoffs.[38]

These same passivities and activities are discernible in reasoning. Most of it happens passively and largely out of sight of the mind's eye. Most of it is down to good machinery. Indeed, it is *always* down to good machinery, but sometimes it is also down to us – to our disciplined, patient and skilled application of the expert routines that knowledge sometimes requires. Think here of the incompleteness of formal arithmetic.

We've already said that our visitors aren't philosophers or logicians of the orthodox sort. They are cognitive anthropologists. This is not Nat's situation. Nat is a philosopher whose dissent is launched from within the very orthodoxies he finds fault with. It might be appropriate for the visitors not to take a philosophical position on knowledge. But Nat can hardly proceed to completion without allowing his epistemic reach to carry *epistemological* implications as well. Nat knows that philosophical theories of knowledge broadly partition into two paradigms, one of them more historically dominant than the other. This first, he calls the Command and Control Model, and the other the Causal Response Model. A typical example of the CC-approach is the JTB model, according to which S knows that *p* if and only if

1. *p* is true.
2. S believes that *p*.
3. S is justified in believing that *p*.

A good example of the CR-approach retains the first two conditions and replaces the third with

3*. S's belief that *p* arises from belief-producing mechanisms that are in good working order and operating in this instance as they should, on good information and in the absence of hostile externalities.[39]

[38] A similar approach is taken, with a good deal more detail and sophistication, in Ernest Sosa, *A Virtue Epistemology: Apt Belief and Reflective Knowledge*, volume 1, Oxford: Oxford University Press, 2007, and volume 2, 2009. See also John Greco, *Achieving Knowledge,* New York: Cambridge University Press, 2010, and *Errors of Reasoning,* chapter 3, notwithstanding modest demurral in chapter 2, p. 53 n. 13.

[39] Since formulating this condition, Daniel Clausén has made me aware of a then-unacknowledged debt to Swedish precursors, writing mainly in analytic jurisprudence. See Per Olof Ekelöf, "Free evaluations of evidence", *Scandanavian Studies in Law,* 8 (1964), 45–66; Edman Martin, "Adding independent pieces of evidence", in B. Hansson, editor, *Modality, Morality and Other Problems of Sense and Nonsense,* pages 180–188, Lund, 1973; and Sören Halldén, "Indiciemekansismer",

We see in this division between the CC and CR models a nice correspondence with the already noted distinction between active-case knowledge, whose attainment is significantly down to us, and passive-case knowledge, whose attainment is largely down to our equipment. Having noticed the statistical dominance of the passive over the active, together with the passive underlay of even the active, Nat concluded that the right base epistemology for a naturalized logic is the CR model, provided that, where indicated, it accommodates the CC model as a proper subtheory for various kinds of knowledge acquisition "up above", in which the knowing agent has an active and self-aware role to play. Think again of the second incompleteness theorem.

The idea that a naturalized *logic* needs a base *epistemology* might strike some readers as implausible – anyhow puzzling. Why would I assume it? The answer is that in the human world reasoning is transacted by cognitive beings, and one of its principal functions is the facilitation of knowledge-seeking and the attainment of epistemic goals. Nat is a philosopher who wants a philosophically tenable account of what makes such reasoning in pursuits of such ends the right way to reason or, as the case may be, the wrong. A logic so designed cannot be judged in the absence of a philosophically convincing understanding of what it is that is facilitated or attained when these goals are in play and properly handled.[40]

Given Nat's naturalistic leanings, this seems much the right choice. Right or not, it is a fateful one. It effects a considerable scrambling of the once-pacific distinction between reasons and causes, and it makes of knowledge a much more causal phenomenon than an intellectually wrought achievement. The same holds for reasoning. By a statistically large measure, our premiss-conclusion reasoning is right when the conclusions we draw are causally induced by belief-producing devices when working as they should. In shorter words still, in the general case you don't have to be smart to reason well; you have to be healthy.

An interesting case in point, and one to which the visiting team paid a lot of attention, is the utter widespreadness of the phenomenon of being got to know things by being told them by others. Both models recognize the phenomenon, but they give it quite different theoretical treatments. The central point of contention is whether *justification* is a general control point for knowledge-transmission, with the CC-theorist voting aye and the CR-theorist nay; that is, nay as a general condition on transmission. The CR-reservation came down to this: If we pay close attention to what happens on the ground, the presence of justificatory involvement is *markedly* less discernible than the levels of recognizable cognitive satisfaction among transmitters and recipients alike. As the visitors came to appreciate, the CC-crowd has spared no effort to attribute the workings of justification even in the absence of its recognizable behavioural presence in the general case. Nat has

Tidskift for Rettsvitenskap, 86 (1973), 55–64. See also Peter Gärdenfors, B. Hansson and Nils-Eric Sahlin, *Evidentiary Value,* Lund: Library of *Theoria*, 1983; and Nils-Eric Sahlin, "How to be 100 % certain 99.5 % of the time". *Journal of Philosophy*, 83 (1986), 91–111.

[40]I am grateful to John Greco for having pressed me on this.

noticed this too. The visitors didn't quite know what to make of it. But Nat knew. What he made of it was that the CC-crowd weren't giving sufficient heed to the respect-for-data rule.[41]

The on-ground data also disclose how much of this causal belief-inducing activity proceeds out of reach of the mind's eye, without notice or attentiveness, without deliberation or overt case-making. As Nat puts it, most of human inference is inference "down below". If this is so, it plays straight into the normativity question. If we wish to remain faithful to the NN-convergence thesis, we will need to derive a naturalistic account of errors of reasoning; for it is to precisely these that the badness of bad reasoning is owed, is it not? Such accounts don't lie idly about. They are not numerously available or free on board. They have to be laboured after. First and foremost perhaps is the reconciliation of our two abundance theses – the abundance of knowledge and the abundance of error. Trailing along is a perfectly natural puzzlement about how, if we're so good at reasoning, why are we so bad at avoiding errors.

Nat has an answer to this which (very sketchily) comes to this: A standing liability for the human knower, no less than the human high-jumper, is that there is only so much he can do. There is only so much that it makes sense for him to want to do – or have the slightest interest in doing. He must learn to live and to set his sights within his cognitive means. As with any design-constrained and resource-limited activity, cognitive success calls for economic alignment of capacity with resource-availability. Nat noted that human individuals are graced with very efficient feedback mechanisms. It is a welcome advantage, enabling a better record at error detection after the fact than before. He concluded from this that it is more economical for a resource-bound individual to correct mistakes after the fact than to avoid them before commission. He also observed that, in case upon case, the human animal is more adept at conclusion-drawing than he is at premiss-selection. Think of the former as errors *of* reasoning, and of the latter as errors *in* reasoning. The dominant fault of premiss-selection is *misinformation*; and it is markedly easier to be misinformed about something than to draw from it the wrong conclusion. On thinking it over, Nat came to the view that

Misinformation and misinference: There are significant global variations in peoples' well-informedness – matching kindred variations in region-to-region levels of ignorance – and yet comparatively uniform performance-levels in human conclusion-drawing.

Nat's naturalizing focus is on premiss-conclusion reasoning, chiefly on the drawing side. If, as it appears, this is something we're uniformly good at, then errors *of* reasoning, when committed at all, must arise from external hostilities or equipment failure. Generally speaking, when we make mistakes of conclusion-drawing, we are dog-tired or strung out, or leveled by a stroke; or the conclusion-drawing equipment

[41]Told-knowledge is discussed at length in chapter 10 of *Errors of Reasoning*. Knowledge occasioned by not being told its opposite is dealt with in chapter 11.

isn't – like the nineteen year old Ford – in quite good enough working order; or we are awash in information-overload. If this is right, something else is also bound to be right. It is that

The thickness of error: Error is a natural phenomenon, with a dominantly descriptive character, but also imbued with what Bernard Williams calls "thickness". Roughly speaking, this means that to find error in a person's performance is to see it as an utterly natural phenomenon but one that is out of joint with the how things are supposed to go.[42]

Since his first course in it, Nat has known that logic from its very beginnings sought for a decent theoretical command of what Aristotle calls *fallacious* reasoning. Since his second course in logic, Nat has also known that the fallacies programme is nowhere in sight in any of the going mainstream logics.[43] It is not hard to see why. We've already said that the modern orthodoxies were built for the relief they promised for metaphysical and epistemological anxieties in the foundations of mathematics. They weren't built for human reasoning, even for when it is transacted fallaciously. This, of course, can't be Nat's own position. Nat wants a logic for real-life premiss-conclusion inference. He wants his logic to solve the normativity problem. He thinks that the correct account of errors of reasoning will be the key to its solution. Surely, one would think, no account of reasoning errors could be complete if it didn't revive and make some headway with the fallacies project.

Nat knows that on the traditional approach fallacies are errors of reasoning having certain distinguishing features. One is that they are attractive, hence inapparent. Another is that they are universal, in the sense that virtually all of us are disposed to commit them with a frequency higher than our error-makings in general. Yet another is their incorrigibility; that is, even after detection and correction, rates of post-diagnostic recidivism are extremely high. Like the rest of us, Nat is also familiar with the traditional list of the fallacies – hasty generalization, *argumentum ad verecundiam* (argument from authority), *argumentum, ad ignorantiam* (arguments from ignorance), and so on. It wasn't long before Nat made a discovery that genuinely surprised him. He saw that a proper regard for the respect for data principle discloses that virtually all the items on the traditional *list* of fallacies have no discernible presence in the instantiation-class of the traditional *concept* of fallacy. Accordingly,

Concept-list misalignment: Virtually none of the fallacies in the traditional list lies in the extension of the predicate "is a fallacy" as traditionally interpreted. Either they aren't errors, or they are not attractively inapparent, or not universal or incorrigible.

[42] *Ethics and the Limits of Philosophy,* Cambridge, MA: Harvard University Press, 1985.

[43] Informal logic is a different story, but it is a story without readers in the mainstream. Recall Burgess' quip that according to the elites "informal logic" is a contradiction in terms. When have we seen a fallacies paper in the *Journal of Symbolic Logic*? The answer is: *Never.*

Here is an example that especially impressed Nat. Nat is on his first trip to Brazil, a country he knows little of, but enough to know that animals called ocelots are resident there. One day, Nat and his Brazilian host are tramping the countryside. "Look", exclaims Luis, "an ocelot!" Nat is surprised. "Good heavens, Luis, I had always imagined ocelots as two-legged, not four." Nat has come to see that, on the basis of a single encounter, ocelots are four-legged. On thinking it over a bit, he also realized that the true generalization "Ocelots are four-legged" is not falsified by the plain and subsequently discovered fact that this other ocelot, Ozzie, a beloved resident of the zoo of his friend Luis' hometown, is three-legged. And in no time at all it became apparent to him that this kind of hasty generalization is, in the human species, as common as dirt (as the saying goes), and that to a quite marked degree the generalizations hastily drawn are actually right, not wrong. Whereupon we have it that

Hasty generalization: Hasty generalization – also called thin-slicing – is not a fallacy in the traditional sense. Indeed, comparatively speaking, it is hardly ever wrong when actually performed.[44]

The obvious question now is whether *anything* instantiates the traditional concept. It is, as I write, an open question in Nat's logic.[45]

The naturalistic turn pulls logic and cognitive science in the opposite direction from the mathematical turn. For three decades and more, the mathematically shaped enquiry has searched out an affiliation with a closer attachment to agent-centred, goal-directed, resource-based, time and action systems. These enrichments are a considerable complication for, whose management more basic formal equipment must be upgraded with new machinery of correspondingly greater complexity, sometimes problematically so. Nat's worry, like my own, about these heavy-equipment upgrades is that the more complex they are to handle, the likelier they are to invite the solace of a new batch of simplifying purpose-built performance norms. Nat isn't opposed to the enlargement of capital assets as such. His reservation about heavy-equipment upgrades is that they leave the normativity problem undealt with. His present inclination is to enrich the logic of premiss-conclusion reasoning with naturalistic assets, especially those of them that improve our grasp of the on-the-ground management of error – its avoidance, its commission, its detection and repair. For it is here that he sees promise of a principled solution to the normativity problem. The empirical turn is an attempt to reshape this enlargement of capital assets, by lightening up on the notion that theory-building is intrinsically theorem-proving.[46]

[44]In this regard, traditional fallacy theorists lag behind some fairly well-established insights of probabilistic models of cognition pertaining to "inductive leaps".

[45]Notwithstanding some answers ventured in *Errors of Reasoning,* chapter 15.

[46]Of course, this is not to overlook that the vigorous mathamatizaion of, say, population biology based on complex computational simulations, as well as various areas of empirical psychology. My view of the matter is entirely straightforward. I will gladly accept any heavy-equipment upgrade

What Nat proposes for logic is what Quine proposed for epistemology. Not everyone thinks that Quine's is a tenable project or, so far, a well executed one. But no one should think that these days naturalized epistemology is anything but a well-accepted part of mainstream epistemology. It is vanishingly unlikely that a friend of naturalized epistemology would dislike Nat's proposal for logic because he distrusts the role of naturalism in philosophy. But he might well dislike it because it lies in the nature of logic not to take well to naturalistic intrusion. This, after all, was Quine's own position. At the core of it all is modern logic's deeply dug-in loathing of psychologism. Epistemology leaves lots of room for psychology, and logic leaves none. Naturalized epistemology may now have found a place in the big leagues, but this is the last thing that one could say of logic. This makes Nat's proposal a radical one for logic if not any longer for epistemology. It also makes Nat's proposal a contentious departure from the still well-favoured normative idealization approach to the social sciences. In a nutshell, what Nat wants is logic's reinstatement of psychologism.

Nat fully acknowledges the efforts of the newer developments in heavy-equipment logics to do better on the score of on-the-ground inference-friendliness. His chief reservations are two. The heavy-equipment technologies don't solve the normativity problem; and bulking up the formal machinery hasn't closed the gap between the logic's theorems and settled practice on the ground. As far as psychology goes, there are weighty autoepistemic considerations to take respectful notice of. If the heavy-equipment crowd thought that there was room for psychology in their projects, they'd have put some in. But they haven't. So they don't.

If the case against the normative presumptions of heavy equipment logics could be made to stand, it carries like consequences for the normative presumptions of the ideal models approach to the social sciences generally. This is getting to be quite a bit of nay-saying. If acquiesced to, all of normatively presumptive science would be put on hold until the normative authority problem is properly sorted out. This separates Nat from virtually all the going traffic in agent-based logics, including logics of probabilistic reasoning, belief-change and decision, epistemic and justification logics, fallacy theory, discourse analysis and normative psychology. Of course, enquiries of every kind are needful of starting points and of assumptions that frame their conceptual spaces. These are assumptions that lend enquiry its procedural and organizational shape. This produces a pair of important consequences for the would-be dissident.

One is that by the very nature of received opinion, there is not much pent up enthusiasm for paradigm-overthrow. The other is that, for want of practice, the orthodoxies' disciples aren't much good at defending them. There is a story making the rounds in which the dialethic logician Richard Routley once challenged the logically more strait-laced David Lewis to prove the classically interpreted law of

---

that a theory is able to pay for, either at the empirical checkout counter or – if only we could find one – the normative checkout counter. More of this can be found in my "Mathematicizing epistemology", *Informal Logic*, 33 (2013), 292–331.

noncontradiction.[47] There was point to his challenge. Routley thought that, when interpreted the right way, the law of noncontradiction did *not* preclude the truth of some select contradictions. So the challenge to Lewis was to show that this couldn't be so. When it came down to it, Lewis didn't bite; he refused to be drawn. He told Routley, in effect, to grow up and stop horsing around. Lewis' was a telling response. It was an outright and unconsidered dismissal.

The point of this little *tableau* is dialectical. Orthodox assumptions carry and are protected by high levels of dialectical *inertia*. So Nat would have been foolish not to have anticipated that his own dissensions might receive scant attention in the high courts of received opinion.[48] On the other hand, Quine and others prevailed to good effect in the aftermath of 1969, against the grain of stiff resistance. So who knows? Perhaps Nat's naturalistic prospects will have brightened forty or so years hence. Notwithstanding our differences of methodological perspective, I say this in the spirit of van Benthem and his colleagues in the closing lines of their Introduction to *Logic in Action*: "And with this much of the five logician actors out of the way, we draw the curtain for this little book – and invite you to enter our world."[49]

Acknowledgements For comments on earlier drafts or discussion of closely related matters, I thank most warmly Johan van Benthem, Franz Berto, Daniel Clausén, Dov Gabbay, John Greco, Maurice Finocchiaro, David Hitchcock, Jaakko Hintikka, Frank Hong, Ralph Johnson, Lorenzo Magnani, Christopher Mole, Adam Morton, Ahti-Veikko Pietarinen, Shahid Rahman, Harvey Siegel, Robert Thomas and Yi Zhao. I also thank two anonymous referees.

Bibliography

Barringer, H., Gabbay, D.M., Woods, J.: Temporal argumentation networks. Argum. Comput. **2–3**, 143–202 (2012a)

Barringer, H., Gabbay, D.M., Woods, J.: Modal argumentation networks. Argum. Comput. **2–3**, 203–227 (2012b)

Burgess, J.: Philosophical Logic, p. 2. Princeton University Press, Princeton (2009)

[47] See my "Wrestling with (and without) dialetheism", The *Australasian Journal of Philosophy*, 83 (2005), 87–102.

[48] In fact, Lewis was not quite true to his word. Instead of just walking away, he asserted that LNC is a truth that neither requires nor admits of proof. This was a mistake, apparently not noticed by Routley. Consider the following:

1. LNC is a true proposition that neither requires nor admits of proof. (Lewis)
2. LNC is true. (from 1, De Morgan and $\wedge$ – elimination)
3. LNC can't be proved. (ditto)
4. But this proves that LNC is true and that it can't be proved. (from 2 and 3, $\wedge$ – introduction)

[49] Precisely; I too would like to see logic back in the world. The actors are Johan van Benthem, Paul Dekker, Jan van Eijck, Maarten de Rijke and Yde Venema, *Logic in Action*, Amsterdam: Institute for Logic, Language and Computation, 2001; p. 5, emphasis added.

Burnyeat, M.F.: The origins of non-deductive inference. In: Barnes, J., Brunschevig, J., Burnyeat, M.F. (eds.) Science and Speculation: Studies in Hellenistic Theory and Practice, pp. 193–238. Cambridge University Press, Cambridge. Reprinted in M.F. Burnyeat, Explorations in Ancient and Modern Philosophy, volume 1, pages 112–151, New York: Cambridge University Press, 2012 (1982)

d'Avila Garcez, A., Gabbay, D.M., Ray, O.,Woods, J.: Abductive reasoning in neural-symbolic systems. Topoi. **26**, 37–49 (2007)

d'Avila Garcez, A., Barringer, H., Gabbay, D.M., Woods, J.: Neuro-fuzzy Argumentation Networks (in press)

Dewey, J.: The Later Works, 17 vols. Southern Illinois Press, Carbondale; vol 12 (1981–1991)

Dorit, B.-O.: Expressive communication and continuity scepticism. J. Philos. **60**, 293–330 (2013)

Edman, M.: Adding independent pieces of evidence. In: Hansson, B. (ed) Modality, Morality and Other Problems of Sense and Nonsense, pp. 180–188. Lund (1973)

Ekelöf, P.O.: Free evaluations of evidence. Scandanavian. Stud. Law. **8**, 45–66 (1964)

Fagin, R., Halpern, J.Y., Moses, Y., Vardi, M.Y.: Reasoning about Knowledge. MIT Press, Cambridge, MA (1995)

Finocchiaro, M.: Arguments about Arguments, pp. 6–7. Cambridge University Press, New York (2005)

Gabbay, D.M., Woods, J.: Formal approaches to practical reasoning: A survey. In: Gabbay, D.M., Johnson, R.H., Ohlbach, H.J., Woods, J. (eds) Handbook of the Logic of Argument and Inference: The Turn Towards the Practical, volume one of Studies in Logic and Practical Reasoning, pp. 449–481. North-Holland, Amsterdam (2002)

Gabbay, D.M., Woods, J.: Agenda Relevance: A Study in Formal Pragmatics and The Reach of Abduction: Insight and Trial, volumes 1 and 2 of A Practical Logic of Cognitive Systems. North-Holland, Amsterdam, 2003 and 2005

Gabbay, D., Woods, J.: Filtration structures and the cut down problem for abduction. In: Peacock, K.A., Irvine, A.D. (eds.) Mistakes of Reason: Essays in Honour of John Woods, pp. 398–417. University of Toronto Press, Toronto (2005); 411–414

Gabbay, D.M.: Equational approach to argument networks. Argum. Comput. **3**, 87–142 (2012)

Gärdenfors, P., Hansson, B., Sahlin, N.-E.: Evidentiary Value. Library of Theoria, Lund (1983)

Gigerenzer, G.: From tools to theories. In: Graumann, C., Gergen, K.J. (eds) Historical Dimensions of Psychology Discourse, pp. 336–359. Cambridge University Press, Cambridge; 339 (1996)

Greco, J.: Achieving Knowledge. Cambridge University Press, New York (2010)

Halldén, S.: Indiciemekanismer. Tidskift for Rettsvitenskap **86**, 55–64 (1973)

Halpern, J., Rêgo, L.: Reasoning about knowledge of unawareness. Tenth International Conference on Principles of Knowledge Representation and Reasoning (2006)

Hamblin, C.L.: Fallacies. Methuen, London (1970)

Harman, G.: Induction. In: Swain, M. (ed.) Induction, Acceptance and Rational Belief. Reidel, Dordrecht (1970). Change in View, Cambridge, MA: MIT Press, 1986; chapter 1

Hintikka, J.: Knowledge and Belief: An Introduction to the Logic of Two Notions. Cornell University Press, Ithaca (1962)

Hintikka, J.: The role of logic in argumentation. Monist **72**, 3–24 (1989)

Hintikka, J., Halonen, I., Mutanen, A.: Interrogative logic as a general theory of reasoning. In: Gabbay, D., Johnson, R., Ohlbach, H.J., Woods, J. (eds) Handbook of the Logic of Argument and Inference: The Turn Towards the Practical, volume one of Studies in Logic and Practical Reasoning, pp. 295–337. North-Holland, Amsterdam (2002)

Makinson, D.: Bridging From Classical to Nonmonotonic Logic, vol. 5 of Topics in Computing. College Publications, London (2005)

Peirce, C.S.: Reasoning and the Logic of Things: The Cambridge Conference Lectures of 1898. In: Keiner, K.L. (ed). Harvard University Press, Cambridge, MA (1992)

Quine, W.V.: Epistemology naturalized. In: Ontological Relativity and Other Essays, pp. 114–138. Columbia University Press, New York (1969)

Sahlin, N.-E.: How to be 100 % certain 99.5 % of the time. J. Philos. **83**, 91–111 (1986)

Sahlin, N.-E., Rabinowicz, W.: The evidentiary value model. In: Gabbay, D.M., Smets, P. (eds.) Handbook of Defeasible Reasoning and Uncertainty Management Systems, vol. 1, pp. 247–265. Kluwer, Dordrecht (1998)

Sosa, E.: A Virtue Epistemology: Apt Belief and Reflective Knowledge, vol. 1. Oxford University Press, Oxford (2007), and vol. 2, 2009

Suppes, P.: Models of data. In: Nagel, E., Suppes, P., Tarski, A. (eds.) Logic, Methodology and Philosophy of Science: Proceedings of the 1960 International Congress, pp. 252–261. Stanford University Press, Palo Alto (1962)

Toulmin, S.: The Uses of Argument. Cambridge University Press, Cambridge, 257; (1958)

van Benthem, J.: Logic and Dynamics of Information and Interaction, p. ix. Cambridge University Press, New York (2011)

van Benthem, J., Dekker, P., van Eijck, J., de Rijke, M., Venema, Y.: Logic in Action. Institute for Logic, Language and Computation, Amsterdam (2001)

van Fraassen, B.: The day of the dolphins: Puzzling over epistemic partnership, In: Peacock, A.D., Irvine, K.A. (eds.) Mistakes of Reason, University of Toronto Press, Toronto, pp. 111–133 (2005)

Williams, B.: Ethics and the Limits of Philosophy. Harvard University Press, Cambridge, MA (1985)

Woods, J.: Mathematizing epistemology. Informal Logic **33**, 292–331 (2013a)

Woods, J.: Errors of Reasoning: Naturalizing the Logic of Inference, volume 45 of Studies in Logic. College Publications, London (2013b); chapters 7 and 8

Woods, J.: Advice on the logic of argument. Revista de Humanidades de Valparaíso, **1**. online version at http://www.revistafilosofiauv.cl/ (2013c)

Woods, J.: Aristotle's Earlier Logic, 2nd revised ed. College Publications, London, to appear in 2015

Woods, J., Irvine, A.: Aristotle's early logic. In: Gabbay, D.M., Woods, J. (eds.) Greek, Indian and Arabic Logic, pp. 27–99, volume 1 of Gabbay and Woods, editors, Handbook of the History of Logic. North-Holland, Amsterdam (2004)

Woods, J., Rosales, A.: Virtuous distortion in model-based science. In: Magnani, L., Carnelli, W. (eds.) Model-Based Reasoning in Science and Technology: Abduction, Logic and Computational Discovery, pp. 3–30. Springer, Berlin (2010)

Chapter 19
Action Models for the Extended Mind

Fernando Soler-Toscano

Abstract Logic has a relevant role in many cognitivist theories by authors like Fodor. Representational theories of mind have space for logical inference. But current trends in cognitivism deal with new topics, like the relevance of the environment in cognitive tasks. The idea of the extended mind focuses on the importance of external resources that can be considered as part of the mind. It seems that logic has nothing to say in these theories. But new advances in dynamic epistemic logic provide tools that allow us to model some of the operations that a cognitive agent makes when interacting with the environment. We do not claim that all aspects of the extended mind thesis can be caught by logical formalisms. But a logical analysis of the epistemic actions related with the cognitive configuration and exploitation of the environment throws light on the novelties of the externalist approaches.

Keywords Epistemic logic • Action models • Extended mind • Rationality • Multi-agent systems

19.1 Introduction

Logic has a relevant role in many cognitivist theories by authors like Fodor (1975). The idea of an intelligent agent that processes information has linked for decades the cognitivist programme with research in Logic and Artificial Intelligence. The AGM model (Alchourrón et al. 1985) is a good example of a logical theory that tries to model changes in an agent's information in accordance with representational theories of mind. The agent has pieces of information and works with them through actions of expansion, contraction, revision, etc.

But current trends in cognitivism deal with new topics, like the relevance of the environment in cognitive tasks. The idea of the extended mind (Clark and Chalmers 1998) focuses on the importance of external resources that can be considered as part

F. Soler-Toscano (✉)
Grupo de Lógica, Lenguaje e Información, Universidad de Sevilla, C/ Camilo José Cela s/n, 41018 Sevilla, Spain
e-mail: fsoler@us.es

J. Redmond et al. (eds.), *Epistemology, Knowledge and the Impact of Interaction*, Logic, Epistemology, and the Unity of Science 38, DOI 10.1007/978-3-319-26506-3_19

of the mind. Now the focus is not in the reasoning agent that handles the information by herself, but in the ability she has to configure and exploit external resources.

It seems that logic has nothing to say in these theories more related with embodiment than with pure reason. But we think that new advances in Dynamic Epistemic Logic (van Ditmarsch et al. 2008) (from now on, DEL) provide tools that allow us to model some of the operations that a cognitive agent makes when interacting with the environment. Some authors call 'epistemic actions' (Kirsh and Maglio 1994) to the interactions that the agent performs with the environment in order to get some information, or simplify some cognitive task. In Sects. 19.3 and 19.4 we use action models (Baltag 1999) to define some epistemic actions with analogous effects to those actions considered by the extended mind thesis.

When modelling the interaction between the agent and her environment the main issue is to solve the omniscience problem. Logical models like AGM produce omniscient agents (Nepomuceno-Fernández et al. 2012). That is, these agents are informed about all logical tautologies and they have all logical consequences of their information. This is a weakness when modelling cognitive agents, but specially for embodied agents, as the main characteristic of these agents is that they are not aware of all the information available to them. They often know only the way for accessing the information, not the information itself. Agents in DEL are usually omniscient, so we need some resource to avoid omniscience. The mechanism we use in this paper is to split the cognitive agent into two logical agents: the agent herself and the environment. We call the environment a *virtual agent* as it represents some (informational) resources of the first agent. We have successfully used this technique to avoid omniscience in the context of security protocols (van Ditmarsch et al. 2012). Though both agents independently omniscient, the non-omniscience of the cognitive agent is obtained by restricting the actions she may perform when interacting with the environment.

The paper is organised as follows. In Sect. 19.2 we present the basic language and semantics that we will use through the paper. Section 19.3 introduces a first epistemic action that will allow the agent to get information from the environment. Section 19.4 is devoted to some of the most specific actions in the external mind thesis: the actions that the agent performs to configure the environment. We consider both introducing into the environment information that was not previously owned by the agent and delegating information previously possessed by the agent. We finish in Sect. 19.5 with some conclusions and lines for future work.

19.2 Language and Semantics

We start by introducing the basic language $\mathscr{L}_P$ that will be used in the rest of the paper. It is a propositional language with modal operators to represent the agent's knowledge and the environment's resources.

Definition 19.1 (Language). Consider a set P of propositions. The *language* $\mathscr{L}_P$ is defined by the following grammar, provided that $p \in P$,

$$\varphi := p \mid \neg\varphi \mid \varphi \wedge \varphi \mid K\varphi \mid E\varphi$$

We read $K\varphi$ as "the agent knows φ" and $E\varphi$ as "the agent knows that the environment has information about φ", that is, the agent knows that using the resources provided by the environment, she can get to know φ or $\neg\varphi$. Other binary connectives can be defined as usual. In the following sections we extend the language by introducing new modalities that allow the agent to interact with the environment.

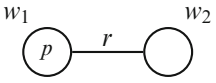
We work with Kripke models with one accessibility relation R for the agent and another one S for the environment.

Definition 19.2 (Models). A *model* for $\mathscr{L}_P$ is a structure $\mathscr{M} = \langle W, R, S, V\rangle$ where,

- W is a countable set of worlds.
- $R, S \subseteq W \times W$ are equivalence relations over W.
- $V : W \mapsto 2^P$, is the valuation function that assigns, to every world $w \in W$, the set $V(w)$ of true propositions in w.

Given $w \in W$, we call $(\mathscr{M}, w)$ a *pointed model*, with w as the distinguished world.

Example 19.1. Look at the following example model $\mathscr{M}_1$:



There are two worlds w_1 and w_2. Propositions included in $V(w)$ are represented inside the circle corresponding to world w. So proposition p is true in w_1 only. The accessibility relations R and S are represented by labelled lines, omitting some trivial links.[1] So, in our example, S contains two equivalence classes $\{w_1\}$ and $\{w_2\}$, but in R there is a single class which contains both worlds.

Now we can interpret the formulas of $\mathscr{L}_P$ in a model. By $(\mathscr{M}, w) \models \varphi$ we express that the formula $\varphi \in \mathscr{L}_P$ is true in the pointed model $(\mathscr{M}, w)$.

Definition 19.3 (Semantic interpretation). Given a pointed model $(\mathscr{M}, w)$, with $\mathscr{M} = \langle W, R, S, V\rangle$, and $p \in P$,

[1] Given that R and S are equivalence relations, as Definition 19.2 states, omitted links can be trivially completed.

$$
\begin{array}{lll}
(\mathscr{M}, w) \models p & \text{iff} & p \in V(w) \\
(\mathscr{M}, w) \models \neg\varphi & \text{iff} & (\mathscr{M}, w) \not\models \varphi \\
(\mathscr{M}, w) \models \varphi \wedge \psi & \text{iff} & (\mathscr{M}, w) \models \varphi \text{ and } (\mathscr{M}, w) \models \psi \\
(\mathscr{M}, w) \models K\varphi & \text{iff} & \text{for all } u \in W, wRu \text{ implies } (\mathscr{M}, u) \models \varphi \\
(\mathscr{M}, w) \models E\varphi & \text{iff} & \text{for all } u \in W, wRu \text{ implies that} \\
 & & \quad \text{for all } v \in W, uSv \text{ implies } (\mathscr{M}, v) \models \varphi, \text{ or} \\
 & & \quad \text{for all } v \in W, uSv \text{ implies } (\mathscr{M}, v) \models \neg\varphi
\end{array}
$$

If $(\mathscr{M}, w) \models \varphi$ for all $w \in W$, we write $\mathscr{M} \models \varphi$ and say that φ is *valid* in $\mathscr{M}$.

We can interpret the operator $E\varphi$ in multi-agent epistemic logic. If a represent our agent and e an agent that models the environment's resources, we can interpret $E\varphi$ as $K_a(K_e\varphi \vee K_e\neg\varphi)$. The operator $E\varphi$ verifies the property

$$\models E\varphi \leftrightarrow E\neg\varphi \tag{19.1}$$

that is, the environment has resources to obtain the truth value of φ iff it has resources to obtain the truth value of $\neg\varphi$.

Example 19.2. Recall model $\mathscr{M}_1$ in Example 19.1. Proposition p is true in w_1 but it is false in w_2, so $(\mathscr{M}_1, w_1) \models p$ and $(\mathscr{M}_1, w_2) \not\models p$. Then, as $w_1 R w_2$, we get $(\mathscr{M}_1, w_1) \not\models Kp$, as there is an accessible world for the agent where p is false. But for S, both worlds are unlinked, so the truth value of p is accessible for the agent by using the environment's resources. Then, $(\mathscr{M}_1, w_1) \models Ep$. Moreover, the formula Ep is also true in w_2, so $\mathscr{M}_1 \models Ep$.

Example 19.2 shows that although the agent does not know some proposition she may know that it is accessible using her external resources. Now we present action models (Baltag 1999), that will allow us to define epistemic actions that the agent may perform to interact with the environment in order to get information. We adapt standard definitions to our semantics with R and S.

Definition 19.4 (Action model). An *action* model is a structure $\mathscr{A} = \langle U, R, S, pre\rangle$, where

- U is a countable set of actions.
- $R, S \subseteq U \times U$ are equivalence relations over U.
- $pre : U \mapsto \mathscr{L}_P$ is the precondition function that assign, for every action $u \in U$, a formula of $\mathscr{L}_P$.

If $u \in U$, $(\mathscr{A}, u)$ is called a *pointed action model* with the distinguished action u.

Actions are executed over models and produce new models. We introduce the formal definition. In the following sections we will provide some examples.

Definition 19.5 (Execution of an action). Given the model $\mathscr{M} = \langle W, R_{\mathscr{M}}, S_{\mathscr{M}}, V\rangle$ and the action model $\mathscr{A} = \langle U, R_{\mathscr{A}}, S_{\mathscr{A}}, pre\rangle$, the *execution* of $\mathscr{A}$ over $\mathscr{M}$ produces a new model $\mathscr{M}' = \langle W', R'_{\mathscr{M}}, S'_{\mathscr{M}}, V'\rangle$, where

- $W' = \{(w,u) \in W \times U \mid (\mathscr{M}, w) \models pre(u)\}$
- $R'_{\mathscr{M}} = \{((w_1,u_1),(w_2,u_2)) \in W' \times W' \mid w_1 R_{\mathscr{M}} w_2, \quad u_1 R_{\mathscr{A}} u_2\}$
- $S'_{\mathscr{M}} = \{((w_1,u_1),(w_2,u_2)) \in W' \times W' \mid w_1 S_{\mathscr{M}} w_2, \quad u_1 S_{\mathscr{A}} u_2\}$
- $V'((w,u)) = V(w)$, for $(w,u) \in W'$.

We call $\mathscr{M} \otimes \mathscr{A}$ the result of executing $\mathscr{A}$ over $\mathscr{M}$.

19.3 Exploiting the Environment

The simplest action that the agent may perform is to access her external resources to get some information. We perform this action, for example, whenever we open an address book. We know that the phone of a friend is there and we go to consult it.

Suppose the agent wants to know whether p is the case. She ignores it but she knows that her external resources can provide her with the information she needs. So she consults those resources and gets to know the truth value of p. We can formalise this process with a simple action model.

Definition 19.6 (Consulting the environment). The action model for *consulting the environment* about φ is

$$\mathscr{A}_{Cons(\varphi)} = \langle \{u_1, u_2\}, I_{\{u_1,u_2\}}, I_{\{u_1,u_2\}}, pre \rangle$$

where $pre(u_1) = \varphi$, $pre(u_2) = \neg\varphi$ and $I_{\{u_1,u_2\}}$ is the identity relation over $\{u_1, u_2\}$.

We introduce in the language the new modality $[Cons(\varphi)]\psi$ meaning that after the agent consults the environment about φ, it is always the case that ψ. Formally,

$$(\mathscr{M}, w) \models [Cons(\varphi)]\psi \text{ iff } (\mathscr{M}, w) \models E\varphi \text{ implies } (\mathscr{M} \otimes \mathscr{A}_{Cons(\varphi)}, (w, u')) \models \psi$$

where

$$u' = \begin{cases} u_1, \text{ if } (\mathscr{M}, w) \models \varphi \\ u_2, \text{ otherwise} \end{cases}$$

The dual modality $\langle Cons(\varphi)\rangle\psi$ is defined as $\neg[Cons(\varphi)]\neg\psi$.

This operation allows the agent to consult the environment about φ without knowing in advance the truth value of φ. The operation $[Cons(\varphi)]$ has the following properties,

$$\models E\varphi \to \langle Cons(\varphi)\rangle \top \tag{19.2}$$

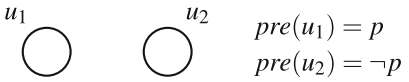
$$\models \langle Cons(\varphi)\rangle\psi \to [Cons(\varphi)]\psi \tag{19.3}$$

$$\models [Cons(\varphi)](K\varphi \vee K\neg\varphi) \qquad \text{(when } \varphi \text{ is propositional)} \tag{19.4}$$

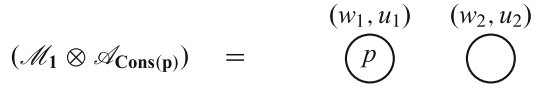
Property (19.2) states that whenever the agent knows that the environment has the information about φ, it is possible for her to consult it. The functionality of the

operation is given by (19.3): if it is possible to perform the operation and obtain ψ, then ψ is a necessary consequence of the operation. Finally, (19.4) indicates the effect of consulting propositional formulas (containing neither E or K): after consulting the environment about φ the agent gets to know the truth value of φ. Example 19.4 shows that (19.4) is not valid for non-propositional formulas.

Example 19.3. Consider again the model $\mathscr{M}_1$ in Example 19.1. As we explained in Example 19.2, the formula p is true in w_1 but the agent doesn't know it. Anyway, she knows that she can get that information by accessing the environment. Following Definition 19.6, the action model for consulting the environment about p, $\mathscr{A}_{Cons(p)}$ is represented in the following picture



when we execute this action in $\mathscr{M}_1$ we get the following model



Now, the agent has learnt from the environment the truth value of p, that is, $(\mathscr{M}_1 \otimes \mathscr{A}_{Cons(p)}) \models Kp \vee K\neg p$.

Example 19.4. Observe that (19.4) is not valid for non-propositional formulas. Look at the example model in the left picture below. In w_1, the agent may consult the environment about $\varphi \equiv p \vee (q \wedge \neg K(p \vee q))$, as the formula $E\varphi$ is true in w_1 (in fact, it is true in all states). But after consulting φ, the model is transformed as the right picture below shows. Now, the agent does not know φ in w_1 because it fails in w_2, and she does not know $\neg\varphi$ as it fails in w_1.

19.4 Configuring the Environment

The agent does not only consult the environment but she also performs some epistemic actions that "alter the world so as to aid and augment cognitive processes such as recognition and search" (Clark and Chalmers 1998). We focus on two actions that augment the environment's resources in two different ways: (1) by incorporating new information that was not previously known by the agent and (2) by delegating some information that the agent had beforehand.

19.4.1 Extending the Environment

Suppose you want to learn about dinosaurs. You do not have any information in your environment about dinosaurs, but you go to the library and borrow a book about prehistorical animals. What are you doing? It is an epistemic action of adding new resources to your environment. You do not know all the details about those resources (you will know them after reading the book) but you know that some of the details you want to know are there.

We can model this operation with an action model that extends the environment with new resources without the agent knowing them.

Definition 19.7 (Extending the environment). The action model for *extending the environment* with φ is

$$\mathscr{A}_{Ext(\varphi)} = \langle \{u_1, u_2\}, \{u_1, u_2\}^2, I_{\{u_1,u_2\}}, pre \rangle$$

where $pre(u_1) = \varphi$ and $pre(u_2) = \neg\varphi$.

We introduce in the language the modality $[Ext(\varphi)]\psi$ to express that after the agent extends the environment with φ, it is always the case that ψ. Formally,

$$(\mathscr{M}, w) \models [Ext(\varphi)]\psi \text{ iff } (\mathscr{M} \otimes \mathscr{A}_{Ext(\varphi)}, (w, u')) \models \psi$$

where

$$u' = \begin{cases} u_1, \text{ if } (\mathscr{M}, w) \models \varphi \\ u_2, \text{ otherwise} \end{cases}$$

The dual modality $\langle Ext(\varphi)\rangle\psi$ is defined as $\neg[Ext(\varphi)]\neg\psi$.

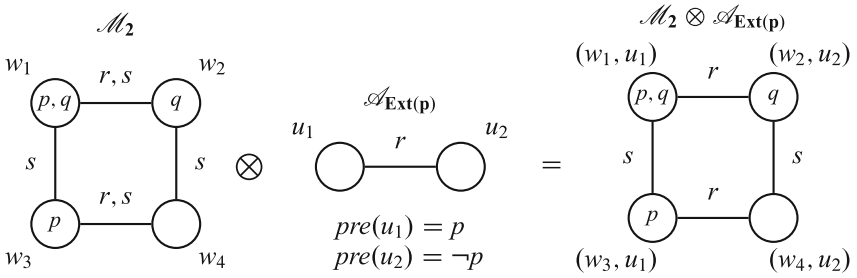
Note that extending the environment with φ does not imply the truth of φ. It's an action to give the agent the information (the truth value, true or false) about φ. Of course, the agent can perform also this operation without knowing the truth value of φ. The idea is that the agent gets the resource φ for the environment without necessarily consulting it. The operation has the following properties,

$$\models \langle Ext(\varphi)\rangle E\varphi \qquad \text{(when } \varphi \text{ is propositional)} \tag{19.5}$$

$$\models \langle Ext(\varphi)\rangle\psi \leftrightarrow [Ext(\varphi)]\psi \tag{19.6}$$

Property (19.5) states that it is always possible to extend the environment with a propositional φ, obtaining $E\varphi$. The functionality and totality of the operation is given by (19.6): the operation can be always performed in a single way.

Example 19.5. The following picture illustrates the effect of extending the environment with p in the model $\mathscr{M}_2$. Like in previous examples we avoid representing some trivial links in R and S.



In $\mathscr{M}_2$ the agent knows the truth value of q, $\mathscr{M}_2 \models Kq \vee K\neg q$, but she doesn't know the truth value of p, $\mathscr{M}_2 \not\models Kp \vee K\neg p$. Moreover, the information about the truth value of p is not in the environment, $\mathscr{M}_2 \not\models Ep$. But in the resulting model $\mathscr{M}_2 \otimes \mathscr{A}_{Ext(p)}$, although the agent continues without knowing the truth value of p, she knows that the environment contains that information, $\mathscr{M}_2 \otimes \mathscr{A}_{Ext(p)} \models \neg Kp \wedge \neg K\neg p \wedge Ep$.

19.4.2 Delegating Resources to the Environment

Suppose you have a dental appointment for next month. It is difficult to keep in mind all the details about the appointment, so you write it in your diary. Probably, you will forget them until some days before the appointment, when you look at the diary page where you wrote the information and recall it. Two epistemic actions have been done. The last one, looking at the diary, can be understood as an action of consulting the environment (Definition 19.6). The first one, writing the information in the diary and forgetting it, is an interesting action of delegating resources to the environment. We make that action whenever we write a note, set an alarm in the cell phone, etc.

Defining the action of delegating information to the environment is more difficult than the previous actions, because it requires the agent to forget some information. We follow the idea of using public assignments and define an *assignment action model* as an action model that can change the truth value of some propositions (van Ditmarsch et al. 2009).

Definition 19.8 (Assignment action model). A *assignment action model* is a structure $\mathscr{R} = \langle S, R, E, V\rangle$ such that

- S is a non-empty set of states.
- $R, E \subseteq S \times S$ are equivalence relations over S.
- $V : S \times P \mapsto \{\top, \bot, I\}$.

where P is the set of propositions.

Symbol $\top$ indicates that a proposition changes to true, $\bot$ to false, and I indicates that the original truth value does not change. Assignment action models are applied to models and change the truth value of propositions, as the following definitions show.

Definition 19.9 (Applying an assignment action). Given the model $\mathscr{M} = \langle W, R_{\mathscr{M}}, S_{\mathscr{M}}, V_{\mathscr{M}}\rangle$ and the assignment action model $\mathscr{R} = \langle U, R_{\mathscr{R}}, S_{\mathscr{R}}, V_{\mathscr{R}}\rangle$, the *execution* of $\mathscr{R}$ over $\mathscr{M}$ produces a new model $\mathscr{M}' = \langle W', R'_{\mathscr{M}}, S'_{\mathscr{M}}, V'_{\mathscr{M}}\rangle$, where

- $W' = W \times U$
- $R'_{\mathscr{M}} = \{((w_1, u_1), (w_2, u_2)) \in W' \times W' \mid w_1 R_{\mathscr{M}} w_2, \quad u_1 R_{\mathscr{R}} u_2\}$
- $S'_{\mathscr{M}} = \{((w_1, u_1), (w_2, u_2)) \in W' \times W' \mid w_1 S_{\mathscr{M}} w_2, \quad u_1 S_{\mathscr{R}} u_2\}$
- $V'_{\mathscr{M}}((w, u)) = (V_{\mathscr{M}}(w) \cap \{p \in P \mid V_{\mathscr{R}}(u, p) = I\}) \cup \{p \in P \mid V_{\mathscr{R}}(u, p) = \top\}$, for $(w, u) \in W'$.

We call $(\mathscr{M} := \mathscr{R})$ the result of executing $\mathscr{R}$ over $\mathscr{M}$.

Note that each state in an assignment action model can change the truth value of every proposition (change to true with $\top$ or false with $\bot$) or leave it unchanged (with I). When applying an assignment action, all the states in the original model are reassigned with all the possibilities in the assignment action. This makes that the number of states can be considerably increased, but we can reduce the resulting model to a bisimilar one (Sangiorgi 2009) as we will see in a later example. Now we can define the action of delegating a proposition to the environment.

Definition 19.10 (Delegating a proposition). The action of *delegating a proposition* $p \in P$ to the environment is an assignment action model

$$Del(p) = \langle \{p_\top, p_\bot\}, \{p_\top, p_\bot\}^2, I_{\{p_\top, p_\bot\}}, V_{Del(p)}\rangle$$

where

$$V_{Del(p)}(s, \varphi) = \begin{cases} \top \text{ if } s = p_\top \text{ and } \varphi = p \\ \bot \text{ if } s = p_\bot \text{ and } \varphi = p \\ I \text{ if } \varphi \in P \setminus \{p\} \end{cases}$$

We extend the language with modalities $[Del(p)]\varphi$ for every $p \in P$, which means that after delegating the proposition p the formula φ is true. Semantically,

$$\begin{aligned} (\mathscr{M}, w) \models [Del(p)]\varphi \text{ iff } (\mathscr{M}, w) \models Kp \vee K\neg p \text{ implies both} \\ (\mathscr{M}, w) \models p \text{ implies } ((\mathscr{M} := Del(p)), (w, p_\top)) \models \varphi, \text{and} \\ (\mathscr{M}, w) \not\models p \text{ implies } ((\mathscr{M} := Del(p)), (w, p_\bot)) \models \varphi \end{aligned}$$

The dual modality $\langle Del(p)\rangle\varphi$ is defined as $\neg[Del(p)]\neg\varphi$.

The operation of delegating a proposition is characterised by the following properties,

$$\models (Kp \vee K\neg p) \leftrightarrow \langle Del(p)\rangle\top \tag{19.7}$$

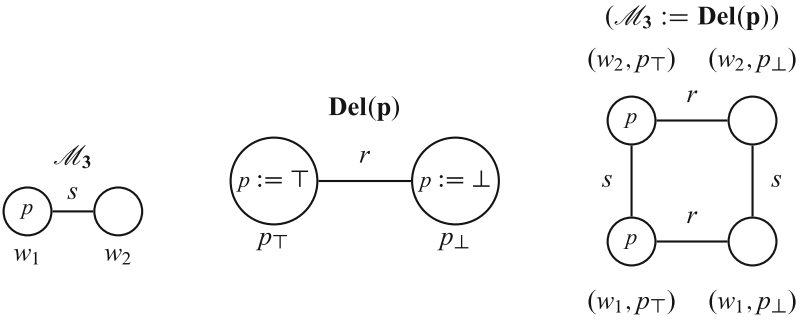
$$\models [Del(p)]Ep \tag{19.8}$$

$$\models [Del(p)](\neg Kp \wedge \neg K\neg p) \tag{19.9}$$

Property (19.7) indicates the condition to perform the operation: the agent must know the truth value of p. The effects of the operation are given by (19.8) and (19.9): the agent knows that the environment gets the information (19.8) and she forgets it (19.9).

Example 19.6. The pictures below depict a model $\mathscr{M}_3$, the action of delegating proposition p and the effect of applying the action to $\mathscr{M}_3$.

The model $(\mathscr{M}_3 := Del(p))$ is equivalent, by bisimulation, to $\mathscr{M}_1$ in Example 19.1. Note the difference of the resulting model with $\mathscr{M}_3$. We have that in $\mathscr{M}_3$, the agent knows the truth value of p, $\mathscr{M}_3 \models Kp \vee K\neg p$ and that the environment doesn't know it, $\mathscr{M}_3 \models \neg Ep$. But after executing $Del(p)$ the agent forgets the value of p, $(\mathscr{M}_3 := Del(p)) \models \neg Kp \wedge \neg K\neg p$, and now the agent knows that the environment knows it, $(\mathscr{M}_3 := Del(p)) \models Ep$.



19.5 Conclusions and Further Work

Current cognitivist theories seem to propose ideas that cannot be approached with formal tools. It is the case with the extended mind thesis. But as we have shown in this paper, by using DEL tools it is possible to define actions that share many characteristics with the action models that a cognitive agent performs when interacting with the environment. We do not claim that all aspects of the extended mind thesis can be caught by logical formalisms. But a logical analysis of the epistemic actions related with the cognitive configuration and mining of the environment throws light on the novelties of the externalist approaches. For example, we have shown that the most sophisticated action is to delegate informational resources to the environment (Definition 19.10), as it requires the agent to forget some information.

The basic idea in our approach has been to consider two agents, the cognitive agent itself and another agent representing the environment. The cognitive agent is aware of the environment's resources and is able to access them. It is possible to extend our proposal by considering several agents that may interchange cognitive resources. In the same way that $E\varphi$ indicates that the agent knows that the environment has resources about φ (Sect. 19.2), DEL provides knowledge operators that allow agents to reason about the information of other agents. That way,

DEL becomes an interesting framework to provide the agents with a *theory of mind* (Premack and Woodruff 1978).

Other action models than those presented in Sects. 19.3 and 19.4 can be considered. In a realistic setting the agent is not aware of all the environment's resources. By using the logic of awareness (Fagin and Halpern 1988) we can define operations to change the formulas that the agent is aware of, in a similar way to Velázquez-Quesada (2010). Then, the agent could access not to all the environment's resources, but only to those she is explicitly informed about.

An interesting application of formal models of cognition is depicted in Galitsky (2002), where the BDI model (Bratman 1987) is used for teaching mental concepts to autistic patients. As the BDI model does not incorporate a theory of mind, predicate logic is used to model the concepts that require to consider beliefs and desires of other people. We think that DEL is a more natural tool to model mental concepts. The possibility of graphical representation of the action models seems an interesting tool for training the recognition of mental attitudes.

Acknowledgements We acknowledge support from the projects FFI2014-56219-P (Ministerio de Economía y Competitividad, Spain) and P10-HUM-5844 (Junta de Andalucía, Spain).

References

Alchourrón, C.E., Gärdenfors, P., Makinson, D.: On the logic of theory change: partial meet contraction and revision functions. J. Symb. Log. **50**(2), 510–530 (1985)

Baltag, A.: A logic of epistemic actions. In: Proceedings of the Workshop on "Foundations and Applications of Collective Agent Based Systems" 11th European Summer School on Logic, Language and Information (ESSLLI'99), Utrecht. Utrecht University (1999)

Bratman, M.E.: Intentions, Plans and Practical Reason. Harvard University Press, Cambridge (1987)

Clark, A., Chalmers, D.J.: The extended mind. Analysis **58**, 7–19 (1998)

Fagin, R., Halpern, J.Y.: Belief, awareness, and limited reasoning. Artif. Intell. **34**(1), 39–76 (1988)

Fodor, J.: The Language of Thought. Harvard University Press, Cambridge (1975)

Galitsky, B.: Extending the BDI model to accelerate the mental development of autistic patients. In: Proceedings of the 2nd International Conference on Development and Learning, Massachusetts Institute of Technology, Cambridge, MA, pp. 82–88 (2002)

Kirsh, D., Maglio, P.: On distinguishing epistemic from pragmatic action. Cogn. Sci. **18**, 513–549 (1994)

Nepomuceno-Fernández, A., Soler-Toscano, F., Velázquez-Quesada, F.R.: Dinámica de la información en agentes no omniscientes. In: Ensayos sobre lógica, lenguaje, mente y ciencia. Alfar (2012)

Premack, D., Woodruff, G.: Does the chimpanzee have a theory of mind? Behav. Brain Sci. **1**(04), 515–526 (1978)

Sangiorgi, D.: On the origins of bisimulation and coinduction. ACM Trans. Program. Lang. Syst. **31**(4), 1–41 (2009)

van Ditmarsch, H., van der Hoek, W., Kooi, B.: Dynamic Epistemic Logic. Springer, Dordrecht (2008)

van Ditmarsch, H., Herzig, A., Lang, J., Marquis, P.: Introspective forgetting. Synthese **169**(2), 405–423 (2009)

van Ditmarsch, H., van Eijck, J., Hernández-Antón, I., Sietsma, F., Simon, S., Soler-Toscano, F.: Modelling cryptographic keys in dynamic epistemic logic with demo. In: Highlights on Practical Applications of Agents and Multi-agent Systems. Volume 156 of Advances in Intelligent and Soft Computing, pp. 155–162. Springer, Heidelberg/NewYork/Dordrecht/London (2012)

Velázquez-Quesada, F.R.: Dynamic epistemic logic for implicit and explicit beliefs. In: Boissier, O., El Fallah Seghrouchni, A., Hassas, S., Maudet, N. (eds.) MALLOW 2010, Lyon. CEUR Workshop Proceedings, vol. 627 (2010)

Chapter 20
Explanatory Reasoning: A Probabilistic Interpretation

Valeriano Iranzo

Abstract This paper deals with inference guided by explanatory considerations – specifically with the prospects for a probabilistic interpretation of it. After pointing out some differences between two sorts of explanatory reasoning – i.e.: abduction and "inference to the best explanation" – in the first section I distinguish two tasks: (a) to discern which explanation is the best one; (b) to assess whether the best explanation deserves to be legitimately believed. In Sect. 20.2 I discuss some recent definitions of explanatory power based on "reduction of uncertainty" (Schupbach and Sprenger 2011; Crupi and Tentori 2012). Even though a probabilistic framework is a promising option here, I will argue that explanatory power so defined is not a convincing characterization of what makes a particular hypothesis better, from an explanatory point of view, that an alternative option. Then, in Sect. 20.3 I will suggest a sufficient condition (rule R1*) as my answer to (a). Regarding (b) I will propose a probabilistic threshold as a minimal condition for entitlement to believe (Sect. 20.4). The rule R1* and the threshold condition are intended as a partial explication of explanatory value (and, consequently, also as a partial explication of "inference to the best explanation").

Keywords Abduction • Bayesianism • Explanatory power • Explanatory reasoning • Inference to the best explanation

In scientific contexts, and also in our daily life, we ask for explanations. Sometimes we have rival explanatory options and we are bound to choose among them. In those situations we use to appeal to their explanatory value and we tend to favour that option which we consider the best one. We are able, then, to assess and compare their respective explanatory merits. This paper deals with explanatory reasoning, that is, inference guided by explanatory considerations, and particularly with the prospects for a probabilistic interpretation of it.

V. Iranzo (✉)
Departamento de Lógica y Filosofía de la Ciencia, Universidad de Valencia – ESPAÑA, Valencia, Spain
e-mail: iranzov@uv.es

J. Redmond et al. (eds.), *Epistemology, Knowledge and the Impact of Interaction*, Logic, Epistemology, and the Unity of Science 38, DOI 10.1007/978-3-319-26506-3_20

After pointing out some differences between two sorts of explanatory reasoning – i.e.: abduction and "inference to the best explanation" (IBE hereafter) – in the first section, I distinguish two tasks: (i) to discern the best explanation; (ii) to assess whether the best explanation deserves to be legitimately believed. Regarding the first task, I discuss some recent definitions of *explanatory power* based on "reduction of uncertainty" (Sect. 20.2). On my view, even though a probabilistic framework is a promising option here, explanatory power so defined cannot give a full characterization of what makes a particular hypothesis better, from an explanatory point of view, that an alternative option. Roughly speaking, explanatory power is just one dimension of explanatory merit. Then, in Sect. 20.3 I will compare alternative strategies in order to detect the best explanation in a set of rival options and I will suggest a sufficient condition (rule R1*) to assess comparisons of explanatory merit.

Nevertheless, discerning that H is the best explanation does not imply that we should endorse it, since H could be a poor explanation, after all. Therefore, to the extent that IBE is an inferential pattern that, by and large, generates justified beliefs, I will suggest a probabilistic threshold as a minimal condition for entitlement to believe (Sect. 20.4). The rule R1* and the threshold condition proposed in Sect. 20.4 are suggested, then, as *a partial explication of explanatory value*. They could also be considered as a partial probabilistic explication of that sort of inference commonly labelled as "inference to the best explanation".

20.1 Two Types of Explanatory Reasoning

Two forms of explanatory reasoning, i.e., abduction and inference to the best explanation, will be compared in this section.

A typical formulation of abduction, taken from Peirce (CP 5.189), goes as follows:

> A surprising fact, F, is observed;
> If G were true, F would be a matter of course,
> Hence, there is reason to suspect that G is true.

F is a particular fact that demands an explanation. G explains F insofar as F is not surprising provided that G is true. G is the "abduced" conclusion that could be considered an explanatory *hypothesis*. Abduction is a very general pattern of reasoning both informative and uncertain. Supposedly, we do not know whether G is true or not. Therefore, G is a *potential* explanation of F: were G false, it would not actually explain F. Although further research is necessary to confirm that G is definitely true, it should be noticed, however, that G may enjoy some plausibility inasmuch as we resort to it instead of preferring some other more weird alternatives. Otherwise G would not even be considered as a potential explanation.

Obviously, if just one potential explanation for F could be found, it could not be any comparison with some other rivals. However, since sometimes there are distinct potential explanations for a particular surprising fact, and it well may occur that they cannot be all true, abduction also requires a comparative step. Given that abduction is guided by explanatory considerations, comparisons are about the

respective explanatory quality of potential explanations. Thus, a posterior selection among the most promising alternatives is mandatory when there are competing explanations.[1]

Now, let us assume an ideal situation where all the potential explanations are ranked according to their explanatory merit. Presumably, the top-ranked hypothesis is the best potential explanation. It should be preferred over the rest as far as explanatory value is concerned. Furthermore, there is some rationale to think it may be true, according to Peirce's quotation. But is that all we can say about it?

Abduction and IBE are closely related. Likewise abduction, IBE can also be counted as "explanatory reasoning". Lorenzo Magnani and Alexander Bird, for instance, equate both (Magnani 2000, 25; Bird 2005, 5). Gerhard Schurz, in its turn, distinguishes several types of abduction, but considers all them as special patterns of IBE (Schurz 2008). There are also some authors who stress the differences between abduction and IBE, but they acknowledge that both sorts of explanatory reasoning are not incompatible (see for instance Campos 2011, 438 and ff.). Furthermore, some AI theorists do not understand abduction working separately from IBE.[2]

However, even though there is a close link between abduction and IBE that goes beyond the fact that they can be considered as particular instances of explanatory reasoning, abduction and IBE cannot be plainly equated. In order to see why it would be worthwhile to discuss Magnani's approach to abduction.

Magnani (2000) considers abduction in a wider context, that is, as a step in a complex abduction-deduction-induction cycle. After detailing the explanatory options, empirical consequences are deduced from them. Then, they are tested by means of induction. From this point of view, even though abduction requires some filter, it is mainly a heuristic procedure for selecting those conjectures which are interesting enough to be subsequently tested. Magnani insists that abduction is not devoid of normative justification, but it should be emphasized that justification comes from empirical testing. An abduced hypothesis may be true or not. It seems more or less plausible but it is not supported, confirmed, . . . , by the evidence it explains. The fact that it is, in principle, an appealing explanation does neither make it true, or more probable.[3]

Perhaps this position is close to Peirce's original proposal about abduction. It is also sympathetic to the account defended by N. H. Hanson fifty years ago (Hanson

[1] All this fits well with Peter Lipton's "two-filter model" (see below). It could be recalled here that perceptual judgments in ordinary conditions were understood by Peirce as "extreme cases of abductive inferences" where the abductive conclusion "comes to us as a flash" (CP, 5.181). Apparently no comparison is done here, although unconscious comparative processes could underlay those judgments. However, provided that these are legitimate examples of abductive reasoning, they are not typical instances.

[2] A good example can be found in Josephson and Josephson (1994) where it is proposed a definition of an "abduction problem" "intended to formalize the notion of best explanation" (p. 160 and ff.).

[3] Minnameier (2004) also discusses abduction and IBE in relation to Peirce's cyclical view of scientific inquiry. He concludes that IBE should not be included in the abductive stage, but in the inductive one.

1965). Nonetheless, the most sanguine contemporary advocates of IBE (G. Harman, P. Lipton, S. Psillos, . . .) think of a more ambitious sort of explanatory reasoning. Harman claimed indeed that, according to IBE, "one infers, from the premise that a given hypothesis would provide a 'better' explanation for the evidence than would any other hypothesis, *to the conclusion that the given hypothesis is true*" (Harman 1965, p. 89; my emphasis). Lipton, another partisan of IBE, characterizes it as follows: "Given our data and our background beliefs, *we infer* what would, if true, provide the best of the competing explanations we can generate for those data (so long as the best is good enough for us to make any inference at all)" (Lipton 2004, p. 56; my emphasis).

Needless to say that both Harman and Lipton allude to Peirce's abduction as an honorable antecedent of IBE. However, in contrast to Peircean or Hansonian abduction, IBE's characteristic features are, firstly, that explanatory value is truth-conducive, and secondly, that the inferential process is completed and a fully justified conclusion/belief is obtained.[4] Both features are related, certainly. The best explanation must be preferred not only because it is better qua explanation than the alternatives, but because it is true (or at least, more probable than the alternatives). To the extent that truth is involved, preference gives way to inference and belief. To sum up, according to this interpretation of explanatory inference epistemic justification is not conferred to it through empirical testing. Rather, the conclusion is legitimately inferred –as a probably true conclusion– *on account of its explanatory value*: "... something would have gone amiss if we thought that the best explanation was *not* reasonably acceptable before it was subjected to Bayesian confirmation." (Psillos 2004, 89–90; see also Mackonis 2013).

It is clear, then, that IBE is not just heuristic abduction, which seems primarily related to the context of discovery. In contrast, IBE clearly assumes a sequential process with three stages: *generation* of potential explanations, *comparison* for discerning which is "the best of the competing explanations", and finally, *inference* to the best one, supposedly a true (or probably true) conclusion. Abduction could be equated to the first and second stages, that is, to the generation and comparison of alternatives –with the proviso that abduction necessarily demands subsequent empirical testing. According to the "two-filter model" of IBE defended by Peter Lipton, selection operates both in the first and in the second stages (Lipton 2004, 67, 148–151). To the extent that IBE incorporates both stages, selection works in it in a similar way to abduction. The distinctive feature of IBE, however, should be found at the third stage where it is assumed that explanatory merit is truth-conducive (Iranzo 2007).

The foregoing discussion shows that abduction and IBE are closely intertwined, but it also suggests deep issues. The first one is "what is explanatory value?". Unless we answer to it, we will not know in which respects the potential explanations should be compared. Secondly, once we elaborate the ranking, the question is

[4]For more details about differences between "Hansonian" and "Harmanian" models of explanatory reasoning, see Paavola (2006).

whether we are entitled to infer the best explanation. Notice that according to the view here defended about abduction and IBE, this is a specific concern for the latter, but not for the former.

In the remaining sections I will cope with both questions from a probabilistic standpoint. It should be remarked, however, that this plan does not take for granted a definite answer to a fundamental question, i.e., what is an explanation. There is no consensus on what is the most appropriate philosophical account on explanation and I will not defend any of the current available options. In particular, I will not assume that the notion of explanation should be analyzed in terms of probability alone. However, it seems reasonable to accept a weaker claim, namely, that the explanatory link between the hypothesis and the evidence explained by it does not necessarily demand a deductive relation (even Hempel accepted this long time ago!). In other words, I will assume that evidence that is not entailed by the hypothesis could be explained by it. Furthermore, while I acknowledge that the nature of this explanatory relation perhaps could not be fully accounted in terms of probability, I also think that it deserves some effort to investigate to what extent explanatory merit can be understood in probabilistic terms. Granted that there may be rival explanatory hypotheses non-deductively related to the same explanandum which differ concerning their explanatory merit, in the next two sections I will explore to what extent their respective explanatory merits depend on the probabilities involved.

20.2 Explanatory Power: Some Critical Remarks

Think about a diagnosis task like inferring diseases from symptoms. The relevant information for setting the search space (i.e.: the lively explanatory options) should take into account those explanatory hypotheses which turn the symptoms into "a matter of course" – recall Peirce's definition of abduction at the outset of Sect. 20.1. We should pay special attention on those hypotheses that make the data *less surprising*. Turning to a probabilistic framework, this is closely related to the *likelihood* of the hypothesis H on the data D to be explained, that is, $p\,(D|H)$. Likelihoods often encapsulate empirical frequencies about how many times H has been followed by D in the past. It is not necessary that D constantly appears when H is the case. Putting the matter in other words, it is not required that D is completely predictable provided that H is true. Rather, D may be more or less expected given H. The point is that hypotheses with low likelihood in respect of D should not be considered as serious explanatory options, since they do not decrease the surprising character of D. In fact, given that a low value for $p\,(D|H)$ entails a high value for $p\,(\neg D|H)$, what can properly be said is that H could be, at most, a good explanation for $\neg D$.

In line with the foregoing considerations, some recent proposals relate "decrease in surprise" to the explanatory merit of a hypothesis. Here are two different measures of "*explanatory power*":

$$Ep_{SS}\,(D,H) = \frac{p\,(H/D) - p\,(H/\neg D)}{p\,(H/D) + p\,(H/\neg D)} \qquad \text{Schupbach and Sprenger(2011)}$$

$$Ep_{CT}\;(D,H) = \begin{cases} \dfrac{p\,(D/H) - p(D)}{1 - p(D)} & \text{if } p\left(D\middle|H\right) \geq p\,(D) \\ \dfrac{p\,(D/H) - p(D)}{p(D)} & \text{if } p\left(D\middle|H\right) < p\,(D) \end{cases} \qquad \text{Crupi and Tentori(2012)}$$

These measures share a common assumption, namely, that the *explanatory power* (*Ep,* hereafter) of a hypothesis depends on "its ability to increase the degree to which we expect the explanandum." (Schupbach and Sprenger 2011, 108)

Although Ep_{SS} and Ep_{CT} are non-equivalent, not even in ordinal terms,[5] they agree on the minimum and maximum values (i.e.: −1 and 1). It is worth noticing that *Ep*-measures obtain these values precisely when there is a maximal reduction in uncertainty, that is, when $p\,(D|H) = 1$ and $p\,(D|H) = 0$, respectively.[6] Thus, if $p\,(D|H) = 1$ – that entails that $p\,(\neg D|H) = 0$– and there is no rival hypothesis in the search space which maximally entails D, then both measures agree that H would be the most powerful explanation for D and also the least powerful one for $\neg D$ (and conversely when $p\,(D|H) = 0$).[7]

Ep-measures do some justice to Peirce's insight on the naturalness of abduced conclusions – see above footnote 1– and there is some progress in elaborating a formal account of that. Besides, *Ep*-measures may be useful in some comparative contexts where inference to truth is not involved. Some of its supporters go beyond, however, and claim that *Ep*-measures give the clue to understand IBE in probabilistic terms (see Schupbach 2011, sect. 5, on the virtues of Ep_{SS}). Yet, even though I agree that the notion of explanatory power may be useful in some particular contexts, I do not think it gives a full account for IBE. My criticism is twofold. Firstly, explanatory power does not suffice since an explanation that enjoys maximal likelihood is not necessarily the best explanation – sometimes it does not

[5] Two different measures M_1 y M_2 are ordinally equivalent just in case, for all H, D, H' and D' it is true that $M_1\,(D,\,H) \geq M_1\,(D',H')$ *iff* $M_2\,(D,\,H) \geq M_2\,(D',\,H')$.

[6] This is obvious for Ep_{CT}. Concerning Ep_{SS}, it should be noticed that $\frac{p(H/D)-p(H/\neg D)}{p(H/D)+p(H/\neg D)}$ equates to $\frac{p(\neg D)p(D/H)-p(D)p(\neg D/H)}{p(\neg D)p(D/H)+p(D)p(\neg D/H)}$ (proof omitted).

[7] For a detailed comparison of both measures, see Crupi and Tentori (2012).

even qualify as a fairly good explanation. Secondly, let us take for granted that the likeliest explanation actually equates to the best explanation in a particular context. Then, if we infer it, according to IBE, it is not guaranteed that we are choosing precisely that explanatory option which is more probable than its rivals according to the available evidence. The point, briefly stated, is that the highest value for $p\ (D|H)$ does not entail the highest value for $p\ (H|D)$. And this is, in principle, an unfortunate consequence for a "likelihoodist" account of IBE, since we should infer as true an explanatory hypothesis that we do not even consider as the most probable among the available options. Let us pause on these objections.

As some historical episodes in science show, counterevidence can be neutralized by means of ad-hoc modifications. An ad-hoc move is made to fit with a recalcitrant anomaly. Usually it makes the disturbing data D a consequence maximally expected, that is $p\ (D|H) = 1$, so the ad-hoc hypothesis is maximally likely –in fact, it was "cooked" to entail the data. But perfect fit is not always related to good explanations. Obtaining maximal likelihood with ad-hoc modifications is, very often, nearly the opposite to obtaining a good, plausible, credible,...., explanation, since ad-hoc hypotheses use to be considered as defective as far as explanatory merit is concerned. Here we have, indeed, a mixed compound of highest explanatory power and low explanatory merit. Thus, it is not only that explanatory power does not suffice to account for the quality of an explanation. The situation is even worse since in some cases favouring maximal likelihood, that is, explanatory power, would lead us astray. It is not clear, then, that "the fact that a hypothesis has more explanatory power over the evidence than any competitor *always provides us with good* (though not necessarily sufficient) *reason*, by the Bayesian's lights, to infer that hypothesis." (Schupbach 2011, 112; my italics). To sum up, to be a powerful explanation, i.e., to enjoy a high value for likelihood, is not a symptom of explanatory merit, even though a low value for explanatory power may indeed be a symptom of low explanatory merit.

We could accept, for the sake of the argument, that sometimes the maximally likely explanation is fairly good. However, maybe it is not the best among the candidates, or, at least, it could not be a good policy to infer it as IBE recommends. Likelihoods are important but Bayesian confirmation theory warns that we should not forget about prior probabilities. If my little niece got up in the morning with headache and fever, we should not be worried about an infectious disease like malaria, for instance, even though it has been reported that malaria causes headache and fever in 99 % of infected patients and bad colds cause this very same symptoms just in 50 % of patients, so to say. Of course, we know that malaria is very rare in Europe so we would say that malaria is not a good explanation in this context, no matter its likelihood. It seems clearly misguided to infer malaria just because of its likelihood provided that we also know that the malaria explanation is much less probable than the alternative. Consequently, we would infer the bad cold explanation as the best explanation for the symptoms and also as an explanation probably true. The formal details go as follows:

Let 'M', 'B' and 'F' stand for 'malaria', 'bad cold', and 'fever'. We know that $p\ (F|B) = 0.5$, $p\ (F|M) = 0.99$. Now, according to Bayes' Theorem,

$$p\left(B\middle|F\right) = \frac{p\,(F/B)\,p(B)}{p(F)} \qquad p\left(M\middle|F\right) = \frac{p\,(F/M)\,p(M)}{p(F)}$$

Therefore, $p\ (B|F) > p\ (M|F)$ when $\frac{p(B)}{p(M)} > \frac{p(F/M)}{p(F/B)}$. So, $p\ (B|F) > p\ (M|F)$ when $\frac{p(B)}{p(M)} > \frac{0.99}{0.5} \approx 2$.

Where do we get the values for $p\ (B)$ and $p\ (M)$? How do we know that $p\ (B) >> p\ (M)$? No expert knowledge is necessary to notice that the odds are favourable to B over M in a ratio superior to 2:1. Presumably the reference class is the people who live in European countries. Let's equate E to "living in a European country". We conjecture that $p\ (B|E) >> p\ (M|E)$, given that malaria is very unusual in Europe, and these values are those of the prior probabilities –i.e.: $p\ (B)$ and $p\ (M)$ – when calculating $p\ (B|F)$ and $p\ (M|F)$, so that the aforementioned condition for satisfying the inequality $p\ (B|F) > p\ (M|F)$ is easily met.[8] Perhaps we should make some further checks before definitively endorsing B. What seems clear, however, is that we should not favour the reduction of the explanandum uncertainty alone since B is, as things stand now, much more probable than M.

The moral of this example is that the explanatory value of a hypothesis also depends on its plausibility leaving aside the particular data to be explained.[9] The relevant information here could be frequencies about the respective rates of malaria and bad colds among people living in Europe, for instance. As a consequence, likelihood is not the only relevant factor when assessing which is the best explanation, that one to be inferred. The contextual plausibility of those alternatives included in the search space –the prior probability, in the Bayesian jargon– must be taken into account.

But, if the relevant information for setting the priors – $p\ (B)$ and $p\ (M)$ – includes known frequencies and these are represented by means of likelihoods, are we not implicitly resorting again to explanatory power? In that case, prior probabilities would be indirectly based on likelihoods, where the explanatory merit of the competing hypotheses is supposedly encapsulated, and not the other way round.

Nevertheless, these considerations do not force us to make any substantial modification on the foregoing comments. Rather, they show that likelihood and explanation may go completely apart. We have just claimed that $p\ (B) \approx p\ (B|E)$ and $p\ (M) \approx p\ (M|E)$. According to Bayes' Theorem, in order to calculate these conditional probabilities we must appeal to $p\ (E|B)$ and $p\ (E|M)$, respectively. Presumably, $p\ (E|B) >> p\ (E|M)$ – the latter is extremely low, insofar as malaria

[8] For Bayesians all probabilities are relative to the background knowledge, so strictly speaking there are no unconditional probabilities. Formalization demands a specific term K at the right of the symbol "|" and $p\ (B)$ and $p\ (M)$ should properly be rephrased as $p\ (B|K)$ and $p\ (M|K)$. Although K is omitted here for ease of exposition, notice that E would be a relevant item included in K.

[9] Weisberg 2009, 129–130, appeals to a different example that highlights the differences about simplicity among the rival hypotheses. However, his point is, again, that our intuitions about comparative explanatory merit are affected by prior probabilities.

affects a few number of Europeans and a huge number of non-Europeans. But, again, it seems distorted to talk here about explanatory merit, even though M is a very powerful explanation of $\neg E$ according to both Ep-measures since $p\,(\neg E|M)$ is very high. The point is that it does not make any sense to claim that being infected by malaria powerfully explains why you do not live in Europe. We do not deny that prior probabilities should be based on background information, if there is any relevant information as statistics or whatever reliable data we could get. But in that case to be based on reported frequencies is not the same as to be based on explanatory merit, unless we beg the question by wrongly equating explanatory power to explanatory merit. Shortly, sometimes explanatory power goes hand to hand with explanatory quality, but some other times not.

It could be argued here that explanatory reasoning is exclusively concerned with comparison of the respective explanatory merits of rival hypotheses. Explanatory reasoning is evaluative and selective, sure, but the goal here is just to discern which option is the best explanation. We are concerned about a particular relation between a hypothesis and the explanandum, and that's all. Therefore, further considerations about the probability of those hypotheses should be entirely put aside.

While this may be a coherent position (Schupbach 2011, sect. 6.1.2), it is not clear that explanatory power so understood –i.e., as a "non-sensitive to priors" factor– is the appropriate notion *to understand IBE* insofar as the alleged truth-conduciveness of explanatory merit is not even considered. Taking for granted that IBE is a sort of explanatory reasoning epistemically sound, it is highly dubious that explanatory reasoning is not concerned with probability (probability of truth, of course). In fact, IBE assumes a link between explanatory value and truth (or high probability) that means that the best explanations are true, or at least, highly probable. This link is at serious risk if we do not take into account the priors.

Yet it could be insisted that, after all, the *most powerful* explanation of the symptom F (fever), although perhaps not the best one, is malaria. Certainly, the most powerful explanation is not necessarily the explanation we should endorse. Furthermore, it must be acknowledged that when there is no reliable information about the priors, the likelihoods may be crucial to assess which is the best explanation. In those situations the most powerful explanation is the option that makes the data less surprising. And surely it is also the best explanation you can get. Nevertheless, it can hardly be accepted that explanatory power exhausts explanatory merit except for very particular examples where prior probabilities cannot be trusted.[10]

To recap, there are strong counterexamples against the equation between maximal likelihood and maximal explanatory merit. Furthermore, even if the maximally likely alternative is the best one in a particular context, it may be not enough good to be inferred, since priors cannot be overlooked. My conclusion is threefold: (i)

[10]Likelihoods are also crucial to discern the best explanation when priors are even. In those situations that explanation which enjoys the highest likelihood would be also the best one. But it is easy to see that Bayes' Theorem entails that it would be the more probable option as well.

reduction of uncertainty is just one component of explanatory merit; (ii) maximizing *Ep* does not always favour explanatory merit (iii) *Ep*-measures do not give a probabilistic rendition of IBE.

20.3 Discerning the Best Explanation

In the foregoing section I claimed that "reduction in uncertainty" does not always increase explanatory value. Initial plausibility –the prior probability– is also relevant. But then, if we are convinced that both likelihoods and priors are necessary to account for explanatory value, a straightforward option would be to equate the explanatory quality of H to the numerator of Bayes' Theorem, i.e.: $p\,(D|H)\,p\,(H)$. The higher the value for it –it ranges from 1 to 0–, the better is the explanation at issue. Given that the denominator of Bayes' Theorem is the same for all the members in the partition, if $p\,(D|H_1)\,p\,(H_1) > p\,(D|H_2)\,p\,(H_2)$, then $p\,(H_1|D) > p\,(H_2|D)$. Therefore, if we take the hypothesis with the highest value for the numerator $p\,(D|H)\,p\,(H)$ as the best explanation, the best explanation is also the most probable one.

A further rationale for favoring this policy is that the comparative stage in IBE is not focused only on discerning the best explanation, as we said before. The goal is to stick at explanations that deserve to be inferred/believed as true (or probable). According to this, it would be disappointing that the best explanation were not also the most probable by the epistemic agent's lights.

I am quite willing to accept that a notion of explanatory merit which allowed an open conflict between "explanatory goodness" and evidence-based conditional probability –that is, $p\,(H_1|D)$–, would scarcely be helpful for a probabilistic account of IBE.[11] Hence, the existence of a perfect match between rankings of explanatory merit and that of probability cannot be taken for granted and demands closer inspection. Couldn't it occur that we identify as the best explanation an option that is below its rivals regarding its conditional probability? Here is a well-known example that apparently supports this claim:

> A cab was involved in a hit and run accident at night. Two cab companies, the Green and the Blue, operate in the city. You are given the following data: (a) 85 % of the cabs in the city are Green and 15 % are Blue; (b) a witness identified the cab as Blue. The court tested the reliability of the witness under the same circumstances that existed on the night of the accident and concluded that the witness correctly identified each one of the two colors 80 % of the time and failed 20 % of the time. (Tversky and Kahneman 1982, 156)

Here we just have two rival hypotheses: H_b (The cab is blue) and H_g (The cab is green). It seems that H_b offers a better explanation of data D (i.e., the witness's testimony that she saw a blue car) than H_g. In fact, if H_b were true, D would not be an amazing fact, while H_g makes D an unexpected event. Nevertheless, H_b is less prob-

[11] J. Weisberg maintains that genuine compatibility between Bayesianism and IBE requires a perfect match between probability and truth, that is, the best explanation is always the most probable hypothesis (Weisberg 2009, 137). The following remarks are intended to show that this is a very demanding condition.

able –conditional to D– than H_g since $p\ (H_b|D) \approx 0.41 < p\ (H_g|D) \approx 0.59$. Intuition gives priority to likelihoods instead of priors in this example. Consequently, the option with the lowest conditional probability would be chosen as the best explanation. Provided that there are just two rival explanations for D, the example shows that the best explanation is not the most probable option. Now, could still be maintained that there is a perfect match between explanatory goodness and probability?

I do not think that the alleged counterexample is so uncontestable. Let us suppose that we describe the same situation in a slightly different way:

> A cab was involved in a hit and run accident at night. Two cab companies, the Green and the Blue, operate in the city. You are given the following data: (a) just one in seven cabs in the city is Blue, and the remaining are Green; (b) a witness identified the cab as Blue. The court tested the reliability of the witness under the same circumstances that existed on the night of the accident and concluded that the witness fails one in five when identifying each of the two colors.

This new description appeals to the same facts, but it emphasizes the failures of the witness. Maybe, it does not suffice to reverse our previous assessment and lead us to consider H_g ("The car is green") instead of H_b as the best explanation. But after reading it we are not so prone as before to conclude that the best explanation of the witness's testimony is that the car was blue. Perhaps now we would suspend judgment about which is the best explanation and ask for more relevant information in spite of the fact that likelihoods and conditional probabilities are the same. It seems, however, that our verdict about the best explanation could be different depending on the description we are given. Shortly, intuitions about explanatory goodness could be affected by the way we account for the situation.

The influence of rhetoric in appraisals of explanatory merit is well-known in some situations – think about lawyers and jurors, for instance. Furthermore, if our perceptive judgments are highly-dependent on the salient features of the situation perceived, why our judgments of explanatory value could not be affected by the salient items of the description?[12] This conclusion should not be a surprise. But the point I want to exploit here is more modest: appearances notwithstanding, the cab-company example does not undermine the view that explanatory merit matches conditional probability.

Certainly, we cannot guarantee that all potential counterexamples could be met in this way. Nonetheless, the problem with the "perfect match" proposal lies elsewhere. The question, on my view, is that it is not clear that a very low likelihood –$p\ (D|H) \approx 0$– could be compensated by a very high prior, since it would be doubtful that in such cases H would even qualify as an explanation of D. Think about a hypothesis H_1 that makes much more expected $\neg D$ than D. Hence, $p\ (\neg D|H_1) >> p\ (D|H_1)$. Let us suppose that $p\ (H_1) > p\ (H_2)$, and $p\ (H_1|D) > p\ (H_2|D)$. According to the perfect match proposal, H_1 should be considered a better explanation of D than H_2. But if $p\ (D|H_2) > p\ (\neg D|H_2)$, it is not clear that H_1 is better than H_2 in

[12] Scientists considerably agree on the "perception" of explanatory goodness, at least on its more general features. This agreement could be a by-product of a highly institutionalized training process that reinforces some heuristics. For details on the cognitive mechanism involved here, see Kuipers (2002).

this context, even though H_1 is more probable. The point is that H_1 could hardly be considered as a legitimate explanation of D. If it explains anything at all, that is not D, but $\neg D$ at most! This means that a threshold condition should also be considered. Unfortunately, it seems difficult to give a precise answer to how much likelihood is necessary for an explanation to be considered a legitimate one.[13] On account of all this, a perfect match between conditional probability and explanatory goodness does not seem a defensible claim. Equating explanatory quality to the numerator of Bayes' Theorem must be consequently discarded.

But there are some other possibilities to compare the respective explanatory merit of two rival explanations. Here is a proposal suggested in Chajewska and Halpern (1997):

R1: H_1 is a better explanation than H_2 if and only if $p\ (D|H_1) > p\ (D|H_2)$ and $p\ (H_1) > p\ (H_2)$

Notice firstly that R1 is not intended *to measure* explanatory merit. After all, comparison does not require measuring. R1 is a rule for selection that can be applied only when likelihoods and priors go in the same direction allowing a conclusive verdict, so it can give us just a partial ordering. According to R1 the best explanation is just the best option among those than can be compared, and not the best explanation among all the available competing options. But partial ordering does not disqualify this proposal. To the contrary, it seems rather realistic to accept that in some contexts there is no conclusive answer about which is the explanatorily superior option in a pair of rival hypotheses –think about my reformulation of the cab-company example. We should accept that those pairs of rivals that do not fulfil the condition stated in R1 cannot be ordered concerning its explanatory respective merits. But no partisan of IBE would be committed to the view that there is always a conclusive answer with respect of which of two rival explanations is better. Then, to acknowledge that total ordering cannot be obtained in some contexts does not raise a challenge for a probabilistic account of IBE.

In addition to this, R1 satisfies the general principle that there is a perfect match between rankings of explanatory merit and those of probabilistic value. In fact, it does not allow any deviation from this principle: if H_1 exceeds H_2 both in likelihood and prior probability, then $p\ (H_1|D) > p\ (H_2|D)$. So, when R1 is satisfied, the partial ordering generated allocates the most probable explanation of those compared at the top rank position. It should also be noticed that R1 is stated as a necessary and sufficient condition so that the relation "__ is better explanation than __" is defined only when R1 is satisfied. There is no point in talking about better or worse explanations if that condition is not fulfilled.

I consider, however, that R1 is too strong. My suggestion here is to replace it by a weaker version:

[13]Perhaps we could assume that in the search space are included only those H_i that satisfy this condition: $p(D|H_i) > p(\neg D|H_i)$. Otherwise, they would not even be considered as putative explanations of D. I will argue in the next section that a similar condition seems reasonable concerning whether *to infer* the best explanation (as a true conclusion that should be believed) or not.

R1*: If $p\,(D|H_1) > p\,(D|H_2)$ and $p\,(H_1) > p\,(H_2)$, then H_1 is a better explanation than H_2

According to R1* it could occur, in principle, that H_1 would be better than H_2 even though those conditions were not fulfilled. Intuitions about examples where likelihoods and priors do not go in the same direction could be, after all, strong, but R1* does not block this possibility, in contrast to R1.

What could be those additional conditions that allow comparisons of explanatory merit when likelihoods and priors do not go on a par? On my view, a probabilistic account of explanatory goodness cannot dispense with an investigation about the factors that underlie our probability assignments.[14] Furthermore, there is a lively debate on the prospects for a Bayesian interpretation of IBE. Some authors think that this project is condemned to fail (B. van Fraassen, S. Psillos), while those labelled as "Bayesian Explanationists" argue that the Bayesian framework is flexible enough to encompass IBE (P. Lipton, S. Okasha, T. McGrew, J. Weisberg, J. Schupbach, . . .).[15] The problem here is whether that probabilistic rendition would be really illuminating. Some authors think that if choosing for the best explanation amounts to favouring the most probable alternative, with no exception at all, it is not clear whether IBE may offer something substantially different from Bayesianism. The risk is "trivializing" IBE just into a sort of informal paraphrase of Bayesianism.[16]

I take it that a fruitful investigation on this topic must assume a remarkable overlapping between verdicts derived from maximizing explanatory merit and those derived from maximizing conditional probability –although this is not the same as a perfect match. My conjecture is that when priors and likelihoods do not go in the same direction, the agent generally tends to consider as the best explanation precisely that option which is the most probable. The point I want to emphasize here is that, usually, the option we discern as the best explanation is also the most probable according to the evidence. Then, in case it deserves to be inferred as a justified belief, we come to believe the most probable alternative *on account of its explanatory merit*.[17] In principle, remarkable overlapping supports the idea that when epistemic agents appeal to IBE they reason *as if they were* Bayesians insofar as they tend to favour those alternatives that would also be favoured by a strict application of Bayesian calculus. Concerning alleged counterexamples against remarkable –albeit not complete– overlapping, my suggestion is either resorting to a neutralizing description –as in the cab company example– or appealing to a

[14] See Mackonis 2013, for a recent proposal.

[15] My particular proposal can be found at Iranzo (2008).

[16] See Schupbach 2011, chap. 4, and Glass 2012, 415 and ff.

[17] Of course, we could eventually discover that the best potential explanation, and also the most probable alternative given the evidence, is false. A justified belief obtained by means of IBE may be false since justified belief $\neq$ true belief, but this is a different question to be dealt with in the next section.

threshold that disqualifies the hypothesis at issue as a putative explanation –in such a way that an extremely low likelihood cannot be compensated by a high prior, for instance. But there is no definite answer for this yet. Meanwhile, it seems interesting to explore to what extent probabilistic notions can illuminate our intuitions about which is the best explanation. In the following section I will turn to a related question, that is, when are we entitled to infer the best explanation.

20.4 Inferring the Best Explanation

Let us take for granted that we know that H_b is the best option in the set of alternatives. Should we infer it? The layman's answer – "Well, it depends on how much good it is" – is a truism. In a Bayesian framework we would rather say "it depends on how much confident you are about it". The point is that the best available explanation of those considered in the search space, could be, after all, really poor. So, what further constraints should be added to ensure that H_b deserves our confidence?

It seems necessary to set a threshold for minimum confidence (i.e., low probability given the data). In order to avoid an arbitrary value, demanding that $p(H_b|D) > 0.5$ seems a reasonable option since it is guaranteed that the alternative inferred is not less probable than its negation.[18] So, I take it that this is a *necessary* condition for confidence in H_b, and consequently, for inferring H_b according to IBE.[19]

This condition, however, does not avoid two disturbing possibilities. The first one has to do with the way we exhaust the range of alternatives. Mathematical probability demands a *partition* of the sample space. That means that the set of H_i must be jointly exhaustive and those H_i must be both consistent and mutually exclusive. Then, unless we are completely sure that the considered options exhaust the sample space, we must keep a place for the "*catch-all hypothesis*", that is, the negation of all those serious candidates $[H^* \equiv \neg(H_1 \text{ U } H_2 \text{ U } ....)]$. But, what should we do if the conditional probabilities of all the serious candidates are low and, consequently, the probability of H^* is higher than 0.5?

To begin with, we would hardly consider H^* as H_b. On the other side, it sounds really weird to maintain that we are rationally forced to infer H^*, given that that is the only option with probability superior to 0.5. That would be a striking misfortune for IBE, since H^* could not be considered a legitimate explanation for D –and it would not be counted as an explanation of D according to any of the extant theories of explanation. If we keep the distinction between the sample space and

[18] Even though we have assumed, *ex-hypothesi*, that H_b is the best explanation, it should be noticed that this condition also guarantees that there is only one best explanation. Thus, if $p(H_b|D) > 0.5$, then $\sum p(H_{i \neq b}|D) < 0.5$; therefore, $p(H_b|D) > p(H_{i \neq b}|D)$.

[19] A different argument for this very same condition was proposed in my (2007).

the *search space* –the latter does not contain all the logical possibilities, but just the live options, so to say–, we should claim that H^* is not even included in the search space. In fact, H^* has no genuine content. All we can do about it is to give a purely negative description. Hence, H^* is no legitimate explanation of D.[20]

Nonetheless, according to our previous threshold condition, no other explanation could be inferred from the set of serious options. Neither of them could reach the appropriate level of confidence, i.e.: being more probable than its negation. That means that we did not find any convincing explanation for D among the H_i considered and inference is blocked since we are not epistemically entitled to believe any of the alternatives. Perhaps we should look for further empirical data in order to get independent support for any of the competing options. Other possibility is to check the explanatory reasoning from the beginning to confirm that we did not miss any relevant alternative – our causal models of the process could be too much simple, perhaps.

As I said at the outset of this paper, the set of rules/conditions developed in Sects. 20.3 and 20.4 are intended as a *partial explication* of explanatory value and the inference guided by it commonly labelled as IBE. They build up a close link between what is good from an explanatory point of view and what is more probable from the agent's particular epistemic position. But the reader might object, and here comes the second disturbing possibility, that the option which is the most probable according to the evidence, could be false. A further alternative, which is not the most probable really, could be true after all and, consequently, $p\ (H_i|D)$ could not be a reliable indicator of H_i's truth.

It should be recalled, firstly, that IBE is a pattern of reasoning both informative and fallible. Fallibility entails that, occasionally, we discover that the explanation previously considered as the best one is false. Therefore, even though by and large the best explanation enjoys higher probability than its rivals, according to the evidence gathered up to now, the world could make it false. Wouldn't that possibility imply that $p\ (H_i|D)$ is not relevant at all to account for IBE's epistemic justification? On my view, this is an overstatement with unacceptable consequences. Provided that the link with truth must be somehow preserved –otherwise, IBE would be devoid of epistemic justification–, IBE would be understood as a sort of cognitive heuristic that allows us to stick at the right option by circumventing the pressure of the evidence. But those who maintained that in these circumstances IBE would still be a justified pattern of reasoning should give some details about the cognitive capacities involved there. The problem is to explain how the agent makes successive lucky guesses about empirical facts, that is, true hypotheses, with no regard to the empirical information.[21]

[20]Incidentally, this is another sort of situations where the most probable alternative given D is not the best explanation of D (see above, Sect. 20.3).

[21]Peirce himself toyed with an alleged "instinctive" ability to stick at the true option. See the discussion in Paavola (2005).

Nevertheless, it must be acknowledged that the normative import of IBE could still be at risk if those explanations we take as the most probable given the evidence would not by and large come out true. Certainly, a probabilistic account of IBE does not solve this problem. It should be added, however, that fallibility is not a specific matter of concern for IBE. It is a constraint for non-deductive reasoning in general. Empirical research, specially about the rate of success in obtaining associated true conclusions in particular situations, is relevant here. That investigation would be on a piece with a naturalized approach to human cognitive resources linked to our ability to elaborate explanations, to discover causal links, ... Computer simulations could also be helpful here.[22] But this issue, fortunately, is beyond the purview of this paper.

References

Bird, A.: Abductive knowledge and Holmesian inference. In: Szabo Gendler, T., Hawthorne, J. (eds.) Oxford Studies in Epistemology, vol. 1, pp. 1–31. Oxford University Press, Oxford (2005)

Campos, D.: On the distinction between Peirce's abduction and Lipton's inference to the best explanation. Synthese **180**, 419–442 (2011)

Chajewska, U., Halpern, J.Y.: Defining explanation in probabilistic systems. In: Proceedings of the 13th conference on uncertainty in AI, pp. 62–71 (1997)

Crupi, V., Tentori, K.: A second look at the logic of explanatory power. Philos. Sci. **79**, 365–385 (2012)

Douven, I.: Inference to the best explanation, dutch books, and inaccuracy minimisation. Philos. Q. **63**, 428–444 (2013)

Glass, D.H.: Inference to the best explanation: Does it track truth? Synthese **185**, 411–427 (2012)

Hanson, N.H.: Patterns of Discovery: An Inquiry into the Conceptual Foundations of Science. Cambridge University Press, Cambridge (1965)

Harman, G.: The inference to the best explanation. Philos. Rev. **74**, 88–95 (1965)

Iranzo, V.: Abduction and inference to the best explanation. Theoria **60**, 339–346 (2007)

Iranzo, V.: Bayesianism and inference to the best explanation. Theoria **61**, 89–106 (2008)

Josephson, J., Josephson, S.: Abductive Inference – Computation, Philosophy, Technology. Cambridge University Press, New York (1994)

Kuipers, T.: Beauty, a road to the truth? Synthese **131**, 291–328 (2002)

Lipton, P.: Inference to the Best Explanation, 2nd edn. Routledge, London (2004)

Mackonis, A.: Inference to the best explanation, coherence and other explanatory virtues. Synthese **190**, 975–995 (2013)

Magnani, L.: Abduction, Reason, and Science: Processes of Discovery and Explanation. Kluwer Academic Publishers, Dordrecht (2000)

Minnameier, G.: Peirce-suit of truth – why inference to the best explanation and abduction ought not to be confused. Erkenntnis **60**(1), 75–105 (2004)

Paavola, S.: Peircean abduction: instinct or inference? Semiotica **153**, 131–154 (2005)

[22] See, for instance, Glass (2012) –where simulations are employed to vindicate a measure of explanatory goodness derived from coherence measures–, and Douven (2013), where the Bayesian standard rule for updating probabilities is disadvantageously compared to an explanationist alternative that gives an extra-bonus to good explanations.

Paavola, S.: Hansonian and Harmanian abduction as models of discovery. Int. Stud. Philos. Sci. **20**(1), 93–108 (2006)
Peirce, C.S. Collected Papers (CP), 8 vols., Hartshorne, C., Weiss, P. (vols. I–VI), and Burks, A.W. (vols. VII–VIII), (eds.) Harvard University Press, Cambridge, MA (1931–1958)
Psillos, S.: Inference to the best explanation and Bayesianism. In: Stadler, F. (ed.) Induction and Deduction in the Sciences, pp. 83–92. Kluwer, Dordrecht (2004)
Schupbach, J. Studies in the Logic of Explanatory Power. Doctoral dissertation, University of Pittsburgh (2011)
Schupbach, J., Sprenger, J.: The logic of explanatory power. Philos. Sci. **78**, 105–127 (2011)
Schurz, G.: Patterns of abduction. Synthese **164**, 201–234 (2008)
Tversky, A., Kahneman, D.: Evidential impact of base rates. In: Kahneman, D., Slovic, P., Tversky, A. (eds.) Judgment under uncertainty: heuristics and biases. Cambridge University Press, Cambridge (1982)
Weisberg, J.: Locating IBE in the Bayesian framework. Synthese **167**, 125–143 (2009)

Chapter 21
The Iconic Moment. Towards a Peircean Theory of Diagrammatic Imagination

Ahti-Veikko Pietarinen and Francesco Bellucci

Abstract Einstein famously said, "Imagination is more important than knowledge". But how to study imagination and how to represent and communicate what the content of imagination may be in the context of scientific discovery? In 1908 Peirce stated that deduction consists of "two sub-stages", logical analysis and mathematical reasoning. Mathematical reasoning is further divisible into "corollarial and theorematic reasoning", the latter concerning an invention of a new icon, or "imaginary object diagram", while the former results from "previous logical analyses and mathematically reasoned conclusions". The iconic moment is clearly stated here, as well as the imaginative character of theorematic reasoning. But translating propositions into a suitable diagrammatic language is also needed: A diagram is for Peirce "a concrete but possibly changing mental image of such a thing as it represents". "A model", he held, "may be employed to aid the imagination; but the essential thing to be performed is the act of imagining" (MS 616, 1906). Peirce had observed that the importance of imagination in scientific investigation is in supplying an inquirer, not with any fiction but, in quite stark contrast to what fiction is, with "an inkling of truth". Since Peirce's limit notion of truth precludes gaining any direct insight into the truth, in rational inquiry the question of what the truth may be or what it could be needs to be tackled by imagination. This imaginative faculty is aided by diagrams which are iconic in nature. The inquirers who imagine the truth "dream of explanations and laws". Imagination becomes a crucial part of the method for attaining truth, that is, of the logic of science and scientific inquiry, so much so that Peirce took it that "next after the passion to learn there is no quality so indispensable to the successful prosecution of science as imagination". In this paper we investigate aspects of scientific reasoning and discovery that seem irreplaceably dependent on a Peircean understanding of imagination, abductive reasoning and diagrammatic representations.

A.-V. Pietarinen (✉)
Chair of Philosophy, Tallinn University of Technology, Tallinn, Estonia

Department of Philosophy, Xiamen University, Xiamen, China
e-mail: ahti-veikko.pietarinen@helsinki.fi

F. Bellucci
Chair of Philosophy, Tallinn University of Technology, Tallinn, Estonia

J. Redmond et al. (eds.), *Epistemology, Knowledge and the Impact of Interaction*,
Logic, Epistemology, and the Unity of Science 38, DOI 10.1007/978-3-319-26506-3_21

Keywords Imagination • Abduction • Peirce • Scientific Reasoning • Discovery • Diagrams

21.1 Introduction

Peirce famously divided scientific inquiry into three stages: first, there is the abductive guess; then follows the deductive derivation of empirically testable consequences and predictions of that guess; finally, the inductive level takes place which subjects the predictions to the test or verification of how well they conform to our experience. This classification became so famous that Richard Feynman repeated it, almost verbatim though without knowing anything about Peirce, in a section of one of his lectures on "The Character of Physical Law": "In general, we look for a new law by the following process. First, we guess it (audience laughter), no, don't laugh, that's the truth. Then we compute the consequences of the guess, to see what, if this is right, if this law that we guess is right, to see what it would imply, and then we compare these computation results to nature or we say compare to experiment or experience, compare it directly with observations to see if it works."[1]

In Bellucci and Pietarinen (2014) we explain how intricate the interlock of these three kinds of reasoning is and how the justification of ampliative forms of reasoning, especially that of abduction (retroduction), partly draws from the most secure, deductive form of reasoning. We also observed that, for the late Peirce, deduction itself consists of two parts. The first part is *logical analysis*; the second is *demonstration*, and includes what Peirce called the corollarial and theorematic types of reasoning. Corollarial reasoning is one in which the conclusions follow trivially from the premises as soon as the premises are known. It is the programmable, algorithmic or mechanistic type of reasoning as for instance exhibited in automated theorem proving. The theorematic reasoning necessitates, in addition, the *creative moment* of introducing and adding some auxiliary individuals or constructions to the proofs. In this sense, the theorematic aspect poses formidable challenges to automatisation of proofs, for instance.

However, besides these two types of proofs that have quite extensively been discussed in the previous literature (Hintikka 1980; Hoffmann 2010; Stjernfelt 2014), there nevertheless is a crucial aspect of deduction that precedes these two, namely *logical analysis*. It concerns not only a detailed analysis of the nature of the problem or question but also the minute development of philosophically and cognitively adequate systems of representation that would capture the essential aspects of the problem, in order to serve future inquiry and to infer the consequences of the hypotheses that the abductive stage has suggested.

The logic of science is in Peirce's view by no means a trivial three-pronged apparatus. Abduction, deduction and induction associate in multiple ways: remarkably,

[1] https://www.youtube.com/watch?v=RXABcv9djQ0 (see also Feynman 1964, 156).

there are abductive and even inductive "moments" in deduction. Abduction occurs in deduction not only at the *demonstrative* theorematic level, as already recognized by several scholars, but also at the *analysis* level, because the operations of analysis, definition, and "representation in a fruitful form" (MS 905) are creative and imaginative components of reasoning just as creative demonstration (theorematic reasoning) is.[2]

The interesting questions are thus of the following kind. Precisely where and how does imagination connect with scientific reasoning? Is imagination based on some pre-linguistic or even non-conventional modes of meaning and representation? Are those modes captured by diagrams of certain sorts? Can those diagrams be logical? What kinds of entities are diagrams, which seem frequently to be resorted to in scientific practice and discovery of something new? Is it here, then, that the methodologies of logic and that of science come to be connected? Is there a link between imagining and drawing diagrams, or perhaps between observing, manipulating and interpreting them? How do these notions show up in Peirce's theory of scientific reasoning and in deductive reasoning?

In this paper, we attempt to answer some of the above questions from the Peircean point of view. In the first place, there is the connection between imagination and abduction. So, in Sect. 21.2, we argue that in explaining abduction, imagination is a superior quality to those of instinct or insight. In Sect. 21.3, we address the role of imagination in science. Section 21.4 presents a sketch of Peirce's theory of imagination, answering the following question: what else, besides the central role imagination has in scientific investigation, can be stated about its nature and content? Section 21.5 argues against the predominantly visual character of scientific imagination.

21.2 Abduction: Instinct, Insight or Imagination?

There are three related but distinct notions at play at the back of Peirce's logical theory of abduction: instinct, insight and imagination. Now mere instinct or instinctive reasoning does not suffice to explain why abductive reasoning functions in the way it does and why it has been successful in the sciences. For instinctive reasoning is

[2]Peirce told his students to possess a "secret" about necessary consequences, which is "a very useful thing to know, although most logicians are entirely ignorant of it. It is that not even the simplest necessary consequence can be drawn except by the aid of Observation, namely, the observation of some feature of something of the nature of a diagram, whether on paper or in the imagination. I draw a distinction between Corollarial consequences and Theorematic consequences. A corollarial consequence is one the truth of which will become evident simply upon attentive observation of a diagram constructed so as to represent the conditions stated in the conclusion. A theorematic consequence is one which only becomes evident after some experiment has been performed upon the diagram, such as the addition to it of parts not necessarily referred to in the statement of the conclusion" (MS 455–6, Lowell Lecture II, 1903).

not logical. Reasoning proceeds from premises to conclusions according to some general reason or principle. But whatever we might mean by "instinct" it fails to refer to any such reason according to which we would be licensed, by voluntary and controlled means of behaviour, to infer certain conclusions from given premises. Instinctive action does not possess a logical form. Resorting to instinct would thus be to admit that no explanation was given at all why we should generate and attune to certain peculiar hypotheses rather than some others, which is exactly what a logic of discovery or, in Peirce's terms, a theory of abduction is supposed to provide.

In some places Peirce suggests that the generation of abductive hypotheses acts like a "flash of insight":

> The abductive suggestion comes to us like a flash. It is an act of insight, although of extremely fallible insight. It is true that the different elements of the hypothesis were in our minds before; but it is the idea of putting together what we had never before dreamed of putting together which flashes the new suggestion before our contemplation. (CP 5.181, 1903)

Some have taken such insight to be a special property in the faculty of the mind that seeks unity and coherence in experience. However, it seems that what actually happens in abduction cannot solely be based on some such singular mental desires of seeking unity in experience. It is true that, according to Peirce, abduction involves, and actually begins with, a "colligation ... of a variety of separately observed facts about the subject of a hypothesis" (CP 5.581, 1898). But that organization and colligation of facts is of the nature of an *inductive* moment, which is involved or embedded in abductive reasoning but not reducible to it.

That colligation does not explain abduction is also seen from the fact that Peirce had learned from Whewell that all reasoning, whether necessary or probable, begins with the colligation of facts. The Greek term for induction – Aristotle's ἐπαγωγή (*An. Pr.* II.23) – means colligation (EP 2, 45, 1898). But also deduction involves colligation. Before the "discoveries" of theorematic reasoning in 1901 and of logical analysis in 1908, Peirce described deduction as composed of colligation, iteration and erasure, and performed through the observation of the results of these operations. In fact, any piece of deductive reasoning whatever always begins with the colligation of the premises. Thus the operation of colligation is found across all the varieties of reasoning and cannot be used to explain the nature of abductive reasoning.

Moreover, it smacks of an unjustifiably psychologistic explanation to resort to needs and desires in explaining the logical workings of the mind. It is not what a pragmatist and logical explanation of those largely hidden steps in abductive reasoning ought to encompass. What we want to understand is how that colligation of facts precisely speaking works, and what are its conceivable effects, not what our singular desires and wishes may be which motivate the colligation. Although sudden acts of insight have repeatedly been reported in scientists' testimonies in relation to the peculiar feelings and experiences they have undergone during those fantastic moments of creative discovery, those acts or flashes of insight do not quite capture the totality of what those special moments consist of, let alone analyze in

the minute detail the complex inferential steps that could have led to such feelings and experiences.

No singular act of sudden flashes of insight is thus sufficient to explain what is going on in the process of colligation of observed facts. That process rather has to do with observing some rational relations, sufficiently compelling and general, that may commonly and invariantly be involved in those facts. Peirce talks about skeleton diagrams and schematic forms: being mental, those skeleton diagrams and forms that we observe in our minds are bound to be indeterminate.

But abduction must also overcome the facts in the sense of reaching beyond them. The logical process of reasoning involved in it needs to proceed beyond the colligation and look further than what any collection of facts can give to us. How to look beyond the facts crucially is what abduction is intended to accomplish. Abduction is that mode of reasoning particularly suited for the cases in which the facts themselves have already largely run out and one must therefore look for some other, collateral means of settling upon some compelling hypotheses in the rational process of scientific guessing that deals with fundamental uncertainty (Pietarinen 2015). At this crucial moment of coming to discover something new, the reasoning no longer can do with the rather mechanical enumerative kind from various particularly observed facts into the colligated wholes of those facts. It has to do with formulating new questions and with looking for some alternative sources of information that could carry the investigation forward, even in the absence of any significant further body of data at one's disposal.

If the insight is, on the other hand, meant to account for what is peculiar to the creative aspect of any discovery, the act-of-insight story fails to take into account the usual preconditions ascribed for creative discovery, such as the ever-so-often tedious but necessary groundwork needed for the insights to arise in the first place, or the equally dull but important consideration of familiar experiences and common sense according to which some insights are to be preferred over some others. Sudden acts of insight are hardly usable in analysing the mysteries of creative discovery, because those acts of insights already consist of those creative moments. In other words, the acts of insights alone are useless in accounting for creative discovery, since those acts readily involve abductive reasoning, namely a creation of those very insights as well as a selection of such insights that would carry the investigation further than would be the case with many other, alternative insights.

What we are left with then is the concept of imagination. Imagination has none of the shortcomings that instinct and insight have. Peirce says that the importance of imagination in scientific investigation is in supplying an inquirer with "an inkling of truth" (CP 1.46, 1896). Since the limit notion of truth precludes gaining any direct insight into the truth, in rational inquiry the question of what the truth may be needs to be tackled by imagination. It is here that the inquirers, blessed with that precious capacity of imagining the truth, will commence the process that "dreams of explanations and laws" (CP 1.48). Thus imagination becomes a crucial part of the method for attaining truth, that is, of the logic of science and scientific inquiry, so much so that Peirce is led to pronounce that "next after the passion to learn there is

no quality so indispensable to the successful prosecution of science as imagination" (CP 1.47).

The most important of the three factors in abductive reasoning is thus imagination. For Peirce, its role in scientific discovery is indispensible. But did Peirce provide a theory of imagination? What else, besides its indisputably central role in scientific investigation, can be stated about its nature and content? We begin by a couple of remarks concerning the role of imagination in science.

21.3 Imagination in Science

Compared to many other notions, imagination has received somewhat less attention in the literature in relation to Peirce's logic of scientific reasoning. One reason may be that Peirce did not really explicate it in so many words. More likely though, an explanation for the silence (but see Tiles 1988 for an exception) lies in the presumed extra-logical or psychological character of imagination, such as its being a mere stimulus for cerebral activities of altogether different sorts, or even in the idea that takes imagination to be on a par with fantasy and fiction. Accept this and one has at once made the notion sound either altogether irrelevant or simply too vague and formless to be of much use in any serious theorising about the logic of science and scientific reasoning.

Such presumptions are highly precarious, however. Already Aristotle, in the third book of his *De Anima*, discusses φαντασία (imagination) as a logical concept, which at the same time is both rational and creative. For Aristotle, all reasoning is imagistic or pictorial in nature, a kind of mental modelling activity. Imagination is "that through which some image comes about for us" (*De An.* III.3, 428a1). He goes as far as to state that "the soul never thinks without a φάντασμα (mental image)" (III.7, 431a 14–17). Other animals as well have sensory imagination, while only man has deliberative imagination: "sensory imagination [...] is present even in the unreasoning animals, while deliberative imagination is present in the reasoning ones" (III.11, 434a). And deliberative imagination is rational imagination, or the ability to imagine different courses of action and their possible results. In this sense, imagination is a condition of appetition, which tends towards an object that is not actually present but only present in the imagination.

For Descartes imagination is one of the four cognitive faculties, along with sense perception, memory, and intellect. But it is Kant who placed imagination at the centre of the description of our cognitive structures: imagination mediates between intuition and understanding in all applications of conceptions to objects of experience. It is a pure faculty, and irreducible to either sensation or intellection. And it is Kant where Peirce starts from.

What imagination is capable of producing seems not only under Peirce's conception but also under those of the eminent scientists' to largely be an opposite category to that of fiction. Take Einstein, for instance, for whom the creations of the mind and scientific imagination are parts of reality – past, present and future:

> If you wish to learn from the theoretical physicist anything about the methods which he uses, I would give you the following piece of advice: Don't listen to his words, examine his achievements. For to the discoverer in that field, the constructions of his imagination appear so necessary and so natural that he is apt to treat them not as the creations of his thoughts but as given realities. (Einstein 1973, 163)

This mindset is echoed by Feynman:

> Our imagination is stretched to the utmost, not as in fiction, to imagine things which are not really there, but just to comprehend those things which *are* there. (Feynman et al. 1964, 127–128, "The Character of Physical Law")

Imagination does not pertain to uncritical, aprioristic metaphysics. There is a logical path from experience and data to theories. Hypotheses do not fall from blind guessing but from the living habits of reasoning in abduction. However, although imagination has an indubitable and important role to play in the sciences of discovery, it is far from clear whether the empirical and cognitive models for imagination that have been proposed, such as those variously using mental and visual models, analogical reasoning, simulation types, blending, perceptual and cognitive theories of thought experiments (Nersessian 2002), or those likening such mental models to neural processes (Thagard 2010), are on equal footing in their attempts to answer the question of the exact and logical nature of imagination, let alone its relation to abductive reasoning.[3]

Nor is scientific imagination unconstrained. It is Feynman's "thinking in a straightjacket". Feynman went as far as to claim that "scientific imagination must be consistent with all else that is known" (Feynman et al. 1964, 20–21). In slightly milder terms, it is in imagination that scientists can check whether something is possible. Peirce's example is "how a body would move upon a vessel itself moving" (CP 6.567); how the established facts at the same time both constrain and direct the creation in the imagination of what is not yet known or brought to the conscious levels of scientific thought. In many suchlike cases, a full-blown experimental study on the relevant issues may not be needed, at least not straight away. A good thought-experiment can serve important scientific and not merely philosophical ends. Examples range from quantum physics (e.g., Schrödinger's cat, EPR "paradox", GHZ experiment) to biology and planetary geosciences (e.g., Levinthal's "paradox", Baker's geomorphological reasoning and his cataclysmic theory of megafloods, see e.g. Baker 1996). The "conceivable practical consequences" that are in the business of pragmaticism to trace out can well be produced in imagination. In fact for Peirce

[3] For example Nersessian (2002, 137) simply dismisses abduction as that mode of reasoning worthy of further consideration in model-based accounts in scientific discovery, on the grounds that it has not been specified what the nature of the underlying processes of abductive inference really is. That is, the charge is that abductive inference does not seem to follow rules. But abduction does have rules, although they are rules of different kind from rules of inference in deductive or inductive arguments, or even from those of reasoning by analogical modes (see Pietarinen and Bellucci 2014 on what Peirce's notion of retroduction/abduction in relation to the other two stages of reasoning seems to amount to).

an experiment – and not only a thought-experiment – is an "operation of thought" (CP 5.420).

What follows from abducted hypotheses is not and need not be the sole affair of deductive types of computation: what the experimental effects and consequences of given hypotheses are is the matter of selection which involves abductive kinds of reasoning, namely those that follow some general and strategic rules such as those of the economy of research. And being in general matters of imagination, there need not be anything directly *sensible* in those statements of conceivable consequences and conditional resolutions to act in certain ways in certain kinds of circumstances. It suffices that the consequences that various hypotheses have are *experienceable*, that is, that they could be produced in scientific imagination capable of *observation*. But the experienceable hypotheses producible in imagination must be those that violate neither the rules of logic nor the laws of nature – or, for that matter, any other facts, laws and constraints already established by previously accepted theories.

> If, besides being a Critical Common-sensist, he is also a pragmaticist, he will further hold that everything in the substance of his beliefs can be represented in the schemata of his imagination; that is to say, in what may be compared to composite photographs of continuous series of modifications of images; these composites being accompanied by conditional resolutions as to conduct (CP 5.516).

What follows from these considerations is for instance that the acts of conceiving and acts of imagining are two quite distinct acts. Conceivable consequences are reproducible in imagination, but not everything that is representable in imagination is translatable to what is conceivable. In a sense, representations in imagination are mental, largely indefinite, imagistic, and predominantly structured by diagrammatic relationships, while what is conceivable has, in addition, somewhat more precision and is structured by a higher degree of conventional meaning, such as what is expressible in counterfactual and subjunctive conditional forms, and what concedes a compelling propensity to act on those forms.

21.4 Peirce on Imagination

Imagination is a recurrent theme in Peirce's discussions of science and scientific inquiry. As supporter of experimental psychology and physiology, Peirce was naturally fascinated by the empirical working of imagination. But as a logician and theorist of science, imagination transcends its psychological nature, and is an element of logical thought proper: "the operation of the imagination [. . .] is most important in all but the lowest kind of thinking" (MS 1114, 1; W4, 43, 1879); "in reasoning of the best kind, an imaginary experiment is performed" (NEM 4, 375, c. 1890); "A pretty wild play of the imagination is, it cannot be doubted, an inevitable and probably even a useful prelude to science proper" (CP 1.235). Not surprisingly for a Kantian thinker as Peirce was, imagination has a crucial role in all reasoning.

A diagram is most often described as a Kantian "schema", and for Kant schemata are the product of pure imagination.

Beginning with his 1885 paper on "The Algebra of Logic", Peirce had claimed that all necessary reasoning is observational, intuitive, imaginative, iconic, and diagrammatic. The object of deduction is always a hypothesis, and a hypothesis is an ideal system or form of relations. Upon such ideal hypothesis, one cannot but reason deductively. But in the first place, such system of relations must, according to Peirce, be either actually perceived or at least *imagined*:

> [. . .] not even the simplest necessary consequence can be drawn except by the aid of *Observation*, namely, the observation of some feature of something of the nature of a diagram, whether on paper or in the imagination. (MS 455–6, 1903)
>
> [. . .] mathematics, which does not undertake to ascertain any matter of fact whatever, but merely posits hypotheses, and traces out their consequences. It is observational, in so far as it makes constructions in the imagination according to abstract precepts, and then observes these imaginary objects, finding in them relations of parts not specified in the precept of construction. This is truly observation, yet certainly in a very peculiar sense; and no other kind of observation would at all answer the purpose of mathematics (CP 1.240).

Thought has to be instantiated in external signs, for "signs are considerably more tangible and overt to examination than ideas otherwise are" (MS 292, 41, 1906). Even when one thinks or calculates within himself, he uses some kind of imaginary diagram to perform that piece of "internal" reasoning. All our thinking "is performed upon signs of some kind or other, either imagined or actually perceived. The best thinking [. . .] is done by experimenting in the imagination upon a diagram or other scheme, and it facilitates the thought to have it before one's eyes" (NEM 1, 122).

Diagrams What is a diagram? Peirce advises us to consider the notion in a quite broad sense:

> In order to expound the truth of my philosopheme that all clearly necessary reasoning is diagrammatic, it will be further requisite that I explain exactly what I mean by a diagram. For I take this word in an extended sense. A Diagram, in my sense, is an Object, whether of sense-perception (more appropriately of vision, but possibly of touch), or of imagination (as ordinarily represented as patched up of pieces of former perception), or of something like inchoate intention. I add this third possibility because I think it must be admitted that a diagram may be of the nature of a Type; that is to say, may be more or less general. For example, a theorem of geometry may be proved by reference to a figure. This figure will generally consist of black lines on a white ground; and if an imagination, as usually represented, is determinate in all respects, it must be of some particular color itself and lie on a ground of a given color. But in the reasoning we pay no attention to the color. We prescind the form from the color; and if images are only reproduced perceptions, we must work with a generalized image, or schema. But I do not know how an image can be generalized and still remain an image, unless, as I say, it be an inchoate intention. Intentions and desires are essentially general, as perceptions and their reproductions are essentially concrete. (MS 293, 1906)

A diagram is an object either of perception (sense perception, vision, touch, hearing etc.) or of imagination. As a material token, the diagram is a singular object, and

thereby determined in all respects. A concrete, sensible image is always necessary in even the most abstract forms of reasoning:

> Now a Diagram is essentially a Sign that is both definite (or not vague) and Determinate (or concrete, in the sense of not being general;) so that something more than vague abstract thought is indispensable in genuine reasoning; and thought that is not brought down to earth by a present sensuous object is, almost if not quite inevitably, both vague and general (MS 633, 8, 1909).

The concrete, empirical image is neither vague nor general: it has a determinate shape, size, color, etc. It is present (i.e., not simply virtual) and sensuous (i.e., given in sense perception). There is a ineliminable element of *direct contact* with the object of reasoning in every inference, for something (be it a chemical sample or a mathematical formula) must in the first place *be given* to perceptual sense.

However, the *intention with which the diagram is constructed* furnishes generality to the empirical representation. The image is to be understood as embodying a *general* concept or an inchoate intention. An image has a purpose, a goal, that is, constructed with intention. The general and formal features of the diagram are those that are teleologically oriented to the purposes of reasoning; the individual and material properties are those that are independent of such purpose. Peirce's solution is Kantian at bottom: the object of investigation in (formal and non-formal) reasoning is a generalized image or *schema*: a concept translated into an "intuitional form" ("Exact Logic", MS 1147, c. 1901) or "intuitional diagram" (MS 17, 8, 1895), into a general symbol which is *schematized* or *diagrammatized* (MS 293, 1907). The diagram has, like Kantian schemata, a "bastard generality" (CP 5.531).

> The purpose of a Diagram is to represent certain relations in such a form that it can be transformed into another form representing other relations involved in those first represented and this transformation can be interpreted in a symbolic statement.
>
> It is necessary that the Diagram should be an Icon in which the inferred relation should be perceived. And it is necessary that it should be in so far General that one sees that accompainments are no part of the Object.
>
> The Diagram is an Interpretant of a Symbol in which the signification of the Symbol becomes a part of the object of the Icon.
>
> No other kind of sign can make a truth evident. For the evident is that which is presented in an image, leaving for the work of the understanding merely the Interpretation of the Image in a Symbol (MS 339, 286r 1906).

Peirce's claim that in a diagram "the signification of the symbol becomes the object of the icon" is really revealing. A symbol is a sign that carries information. Any proposition does so; any term or predicate does so, at least virtually; any argument does, and in a peculiar way (carries information that in its turn will become a source of further information). An icon, on the contrary, is a sign "from which information can be derived" (MS 478, 51–57, 1903). An icon represents the information contained in the symbol in such a way as to render further information derivable from it. In traditional terms, the Icon *denotes* what the Symbol *connotes*.[4] Take

[4]By hypostatic abstraction we convert a term that connotes (a predicate) into one that denotes (a subject). We transfer matter from the signification to the denotation or, as Peirce sometimes

the symbol or concept of triangle. It connotes or implies certain characters (those contained in its definition). By making these characters the object of an icon, that is, in representing them in *an image* instead of simply *thinking* of them, we are *forced* to express other characters not implied in the definition (e.g., that certain relations between the angles subsist). This is why the denotation of the icon is not exhausted by the connotation of the symbol. The icon denotes *more* than the symbol connotes. In representing the signification of the symbol, the icon automatically represents other information not explicitly contained in that signification: the remaining "part" of the object of the icon will be the information that can be derived from the former part. (An icon makes explicit what in the symbol was only implicit.) This is, in semiotic terms, the reason why deduction is informative or, in Kantian terms, synthetic. Deduction is synthetic because it *constructs* its objects, or *schematizes* its concepts: "the Iconic Diagram and its Initial Symbolic Interpretant taken together constitute what we shall not too much wrench Kant's term in calling a *Schema*, which is on the one side an object capable of being observed while on the other side it is General" (MS 293, 1906).

Kant's Schemas In Kant's work the schema is the product of imagination (*Einbildungskraft*, KrV, A 140, B 179). What is brought under the conception is not a representation in general (like in general logic), but the pure synthesis of representations (KrV, A 78–79, B 104). In the first edition of the first *Critique* – which Peirce praised more than the second – Kant makes imagination a third faculty, distinct from sensibility and understanding. In the second, it is affiliated to the understanding (KrV, B 153). "*Imagination* is the faculty for representing an object even without its presence in intuition" (KrV B 151), Kant states. Imagination is the source of the synthesis of manifold, while the understanding is the *representation* of such synthesis, or the bringing of it under rules (understanding is the faculty of rules, KrV A 132, B 171). This synthesis of imagination is *a priori* and is also called figurative (*synthesis speciosa*), and is distinct from purely intellectual synthesis (*synthesis intellectualis*) (KrV B 151). Now, imagination makes a synthesis of representations given in intuitions, and thus in sense perception (for we do not have "intellectual intuitions"). But at the same time, the working of imagination is *conceptually* driven (it is a product of spontaneity, B 151): it is *in accordance with the rule* that the manifold given in intuition is united in a schema or figure. Kant distinguishes therefore *productive* from merely *reproductive* imagination. Productive imagination is the *exhibitio originaria* of the object and prior to the experience of the object, while reproductive imagination is the reproduction of a previous experience of the object. The latter presupposes the working of the former, since any empirical synthesis presupposes a pure synthesis as its condition of possibility. Pure imagination determines *a priori* the intuition of the object.

says, from the interpretant to the object. Hypostatic abstraction is not just the principle engine of mathematical reasoning. Hypostatic abstraction is a very primordial ingredient of every form of thinking whatsoever.

For Peirce and in exactly the same sense the symbol *directs* the construction of the icon. The icon-imagination is the Kantian synthesis of manifold of intuition in an image, while the symbol-understanding is the *representation* of such synthesis, or the bringing of it under rules. These two operations constitute diagrammatic reasoning. But the icon-imagination that represents according to some a priori, symbolic rules cannot be the merely *reproductive* operation of bringing to consciousness past sense-perceptions. The icon-imagination, and the iconic-imaginative moment in reasoning depend on the possibility of *directing* the construction of a perceptual experience. This is unmistakably the task of what Kant called *reine Einbildungskraft* (pure imagination). Peirce describes it as an *act*:

> The word diagram is here used in the peculiar sense of a concrete but possibly changing mental image of such a thing as it represents. A drawing or model may be employed to aid the imagination; but the essential thing to be performed is the act of imagining. (MS 616, c. 1906)

The *act* of imagination *instantiates* the concept: it offers to the senses an empirical image that embodies the concept (Peirce's inchoate intention). The act has of course to materialize somehow its instantiated concept either in outward or inner perceptions, that is, either in empirical intuition or in empirical imagination. But the *act* itself cannot be reduced to anything empirical. Rather, any empirical instantiation of a concept, any empirical image that represents something, presupposes an *a priori* capacity of *directing* the construction of the image. In Kant's jargon, the manifold of intuition is synthetized in accordance with the concept. In Peirce, the diagram is constructed in accordance with the meaning of a symbol.

The heart of diagrammatic reasoning lies for Peirce in what Kant had called a *synthesis figurata* or *speciosa*. The act of imagining itself is independent of the empirical image met with in experience. In order to explain the idea that the product of an act of pure imagination is not an empirical image but a schematic or general image, Peirce uses the fortunate expression "form of a relation".[5] Even in the most concrete scientific experiments, that which is in question is never the individual object of investigation, but the form of a relation that it embodies: "the object of the chemist's research, that upon which he experiments, and to which the question he puts to Nature relates, is the Molecular Structure, which in all his samples has as complete an identity as it is in the nature of Molecular Structure ever to possess" (CP 4.530, 1906). The molecular structure is the form of a relation

[5]Hookway (2010) has argued that Peirce's insistence on the idea of a 'form of relation' suggests a position akin to structuralism in the philosophy of mathematics. Pietarinen (2010a) argues against that view on a number of counts, including structuralism's neglect of (i) experimentation and observation on the diagrammatic forms of relations, (ii) the kinds of reality of objects in structuralism which are all-important in Peirce's theory of signs and philosophy of mathematics, (iii) the continuity of forms, as well as (iv) the essentially hypothetical and modal notions that characterise mathematical assertions. Pietarinen (2014) suggest a closer alliance of Peirce's pragmatist philosophy of mathematics with that of modal-structuralism, although that, too, suffers from nominalism with respect to the semantics of its key objects, namely those of possible worlds (see also Pietarinen 2005, 2011).

which is identical in all samples experimented upon, and is the proper object of the chemist's research. Likewise, in optics, the equation $1/f[1] + 1/f[2] = 1/f[o]$ represents the form of relation "between the three focal distances that these letters denote" (*ibid.*). The proper object of chemical or optical analysis is not the actual, concrete sample or object, but a general feature of it, that is, a complexus of relations and interdependencies among objects and parts of objects. Any kind of reasoning, both scientific and everyday reasoning, has as its object a *system* of relations that has a certain *form*. Any inquiry is, in this sense, a formal inquiry into formal properties. Examples of this kind of pure or formal imagination both in mathematics and in the sciences are plentiful.

Theoric Imagination In mathematics, pure imagination is also called by Peirce "theoric imagination": "In all demonstrations care must be taken to make the theoric imagination as characteristic of the case in hand as possible" (MS 201, 1908); "the plan of a demonstration involves, in some form, the theoric imagination that is it to employ" (*ibid.*). In mathematics, theoric imagination is what allows the mathematician to take a particular logical step in the demonstration that Peirce calls "theoric step". Theoric steps are non-mechanical and non-trivial operations and transformations employed in the demonstration of a theorem. Not all deductive reasoning can be explained by pure "syllogistic" or "corollarial" proceeding. In our analysis of deductive reasoning we must leave room for such inventive steps, because "there is *some* theoric reasoning, something unmechanical, in the business of mathematics" (MS 201, 81, 1908). Theoric steps are not, strictly speaking, deductive: "This part of the theorematic procedure, I will call theôric reasoning. It is very plainly allied to retroduction, from which it only differs as far as I now see in being indisputable" (MS 754, quoted in Hoffmann 2010, 590). The theoric imagination required to demonstrate non-trivial theorems is "plainly allied" to aspects of abductive reasoning. According to Peirce, it is in virtue of an abductive insight that the mathematician can take a peculiar theoric step in demonstration. Often, Peirce argues, the theoric step consists in "looking at facts from a novel point of view" (MS 318, 50). Peirce's favourite example is a theorem of projective geometry that goes under the name of "Ten-points theorem". The theorem is proved by seeing the two-dimensional diagram as three-dimensional, that is, by seeing it as a representation in perspective. According to Peirce, "Everything is corollarial except the single idea that the plane figure is a projection of a figure in three dimensional space. That is certainly not corollarial, since there is nothing in the problem to suggest it, — no reference to a third dimension" (MS 318, 53, 1907; cf. Hoffmann 2010). Interestingly, mathematics requires "an imagination which would be poetical were it not so vividly detailed" (MS 201, 81, 1908). Of course mathematical procedures are grounded upon an accumulating body of mathematical knowledge and past interpretations. But the peculiar imaginative act of the mathematician consists precisely in bringing this body of knowledge to bear on a new perspective that is in part foreign to that knowledge:

> The whole difficulty of mathematics is in imagining from what point of view to consider the facts, and how to bring his store of previous interpretations to bear upon a new one. (MS 662, 7, 1910)

Theoric imagination is the middle term between the body of mathematical knowledge and the demonstration of a new theorem upon the basis of that knowledge. As rightly emphasized by Campos (2009, 139), imagination has a double role: it has an "originative function" in that it is at the origin of hypotheses, but it also has a "transformative function" in that it manipulates and experiments upon the initial hypothesis in order to discover new truths about the object of the hypothesis. It is especially in connection with the "originative function" of imagination that mathematics is ever so often abductive in its mode of proceeding. As a matter of fact, Peirce pronounces, all great hypotheses of mathematics come to us through abduction (MS 754, 6, 1907).

Theoric imagination is perhaps even the most crucial in abductive reasoning, as we have argued in Sect. 21.2 above. According to Peirce, the "greatest piece of Retroductive reasoning ever performed" (CP 1.74) had been Kepler's discovery of Mars's orbits, each logical step of which is recorded in the *Astronomia Nova* (1609). Kepler's abductive reasoning involved diagrammatic experimentation:

> His admirable method of thinking consisted in forming in his mind a diagrammatic or outline representation of the entangled state of things before him, omitting all that was accidental, retaining all that was essential, observing suggestive relations between the parts of the diagram, performing diverse experiments upon it, or upon the natural objects, and noting the results. (W 8, 290)

Such a method of abductive, diagrammatic experimentation cries for "imagination":

> The first quality required for this process, the first element of high reasoning power, is evidently imagination; and Keppler's fecund imagination strikes every reader. But "imagination" is an ocean-broad term,—almost meaningless, so many and so diverse are its species. What kind of an imagination is required to form a mental diagram of a complicated state of facts? Not that poet-imagination that "bodies forth the forms of things unknowne," but a docile imagination, quick to take Dame Nature's hints. The poet-imagination riots in ornaments and accessories; a Keppler's makes the clothing and the flesh drop off, and the apparition of the naked skeleton of truth to stand revealed before him. (*ibid.*)

It is not the poet's imagination that is required in diagrammatic abduction. Artistic imagination reigns free and dreams of "opportunities to gain" (CP 1.48). Rather, what is required is imagination that only focuses on the prescinded, salient and skeletonized features of the phenomenon and leaves aside all irrelevant "clothing".

The object of such "docile", "theoric" and "pure" imagination is not, properly speaking, an image, which is rather the product of empirical imagination. The product of pure imagination is what Kant called a schema and Peirce a diagram: all scientific imagination is diagrammatic, because all reasoning is so, directly or indirectly (MS 293, 1907). Notice that also the converse is true: just like Wittgenstein stated that all pictures are also logical pictures, while not all pictures are chromatic pictures, so did Peirce that all diagrams are logical, and that all diagrammatic thinking is *scientific* thinking ("logical" in Peirce's sense), while not all diagrammatic thinking is, for example, *visual* thinking.

21.5 Pure Imagination not Predominantly Visual

Peirce's talking of "composite photographs", "images", "observation", and the like may nevertheless suggest that what is at issue here is really matter of perception, especially visual perception. However, logical diagrams are just proxies for that which, in order to be communicable, would ultimately need to be stated in generally conceivable terms, such as in linguistic, conventional or mathematical terms. Thus the general concept of diagrams, largely concerned with objects that are mental and imaginary in content, are means to an end and not the final products of knowing. What the common talk about mental images, imagery or pictures is intended to capture does not therefore cover well what is special and salient in diagrams as they are used to aid scientific discovery.

Maybe something like this contingent and weakly understood relationship between diagrams and linguistic means of expression was what Einstein meant when he lamented that

> the words or the language, as they are written or spoken, do not seem to play any role in my mechanism of thought. The psychical entities which seem to serve as elements in thought are certain signs and more or less clear images which can be 'voluntarily' reproduced and combined. .... This combinatory play seems to be the essential feature in productive thought before there is any connection with logical construction in words or other kinds of signs which can be communicated to others . . . The above mentioned elements are, in my case, of visual and some of muscular type. Conventional words or other signs have to be sought for laboriously only in a secondary stage, when the mentioned associative play is sufficiently established and can be reproduced at will. . . . The play with the mentioned elements is aimed to be analogous to certain logical conceptions one is searching for. (Einstein, quoted in Hadamard 1949, 142–143).

It would be quite misleading to consider these testimonies as articulations of peculiarly visual methods of thinking in individual scientists, however. Visuality is far too one-sided a notion and not in fact even a necessary requirement for something to be a diagrammatic representation (including diagrammatic in the sense of logical languages, Pietarinen 2010b). Classifying scientific geniuses into two, on the one hand the linguistic types, and on the other the non-linguistic or visual thinkers, is likewise bound to result in oversimplifications. It might even be a false dichotomy. Logical diagrams must be grounded on systems of conventions, while the category of diagrams is by no means exhausted by visually representable ones. To wit, Einstein alludes in the above passage to muscular feelings, and interestingly enough this is not only in the spirit but even in the letter of what Peirce had written:

> We form in the imagination some sort of diagrammatic, that is, iconic, representation of the facts, as skeletonized as possible. The impression of the present writer is that with ordinary persons this is always *a visual image, or mixed visual and muscular*; but this is an opinion not founded on any systematic examination. If visual, it will either be geometrical, that is, such that familiar spatial relations stand for the relations asserted in the premisses, or it will be algebraical, where the relations are expressed by objects which are imagined to be subject to certain rules, whether conventional or experiential. This diagram, which has been constructed to represent intuitively or semi-intuitively the same relations which are abstractly expressed in the premisses, is then observed, and a hypothesis suggests itself that

> there is a certain relation between some of its parts – or perhaps this hypothesis had already been suggested. In order to test this, various experiments are made upon the diagram, which is changed in various ways (CP 2.778).

Such sentiments on the embodied yet not necessarily visual nature of skeleton diagrams that can represent scientific facts were shared by Feynman in his own playful way, for instance when he was attempting to perceive, sometimes even by listening, the possible movements and trajectories of electrons. One can now also begin to see what Feynman could have meant when he intimidated the interviewer who kept on pressing him on "so what you really see are visual pictures?" To this Feynman replied: "You keep on repeating that. What I am really trying to do is bring birth to clarity, which is really a half-assedly thought-out-pictorial semi-vision thing. I would see the jiggle-jiggle-jiggle or the wiggle of the path. Even now when I talk about the influence functional, I see the coupling and I take this turn—like as if there was a big bag of stuff—and try to collect it in away and to push it" (Glock 1992, 225).

Diagrammatic representations are involved at various stages of the process of inquiry, including abduction, logical analysis and theorematic reasoning. Although we may hold those representations to be predominantly mental, albeit not necessarily visual constructions, they can be put into the format that makes the objectivity of their contents intersubjectively testable: "diagram", it is worth repeating, is for Peirce "a concrete but possibly changing mental image of such a thing as it represents. A drawing or model may be employed to aid the imagination; *but the essential thing to be performed is the act of imagining*" (MS 616, 1906, our emphasis). Models aid imagination but are subservient to it.

In a similar vein, we feel that the question of the nature and justification of different types of models in science has been somewhat overplayed in recent philosophy of science, while much less effort has been levied on the question of the discovery of such models. The prevalent talk on idealized, abstract or approximate nature of models tends to take the key issues to a direction away from the objects of reality rather than closer to them. For how are the models discovered in the first place? How do iconic forms of reasoning, as congenial parts of the process of creation of hypotheses, aid the discovery of models? Note that models can be very concrete forms yet have generality just as diagrams do, as both are constructed according to the general rules and precepts one finds governing a multiplicity of phenomena. Models represent conceivable states of affairs that can be stated in hypothetical conditional forms. They thus have the status of the hypotheses, but imagination is needed to produce them. Frigg and Hartmann (2012) have stated that "no theory of iconicity for models has been formulated yet" – but in our interpretation of Peirce's comments above we have the crucial elements of what such an iconic moment in hypothesis generation would consist. In those comments we thus find not only an account of scientific representations in terms of his three-place sign relation but also of models as icons as the latter occur in scientific discovery.

But the mere acts of imagining and their products would happen in vain unless there be ways of translating the representations of phenomena into suitable *logical* representations by which one is to communicate the contents of those acts and to deduce testable conceivable consequences. But these outward representations in a concrete and intersubjectively communicable and testable medium are equally diagrammatic, as they are constructed according to the very same rules that apply to the concrete mental representations that the acts of imagining create.

21.6 Conclusions

The value of diagrams in imagination is thus seen to be in such forms that provide optimal conditions for the facilitation of scientific reasoning beyond currently available data. Their value is therefore not reducible to merely heuristic devices or placeholders for modal-type thought-experiments. When Norton (1996) claims that thought-experiments are simply disguised inductive, deductive or enthymematic arguments, he overlooks the real possibility that in thought-experiments we in fact follow the rules of abductive reasoning. Moreover, diagrams need not be visual even though they are seen by the Mind's Eye. They are cognitive and analytic instruments embodying what is essential in the reasoning of the mind and in the laws of its behaviour. Their peculiarity is in that they represent reasoning "outward" as well as indicate to the mind "inward" what the nature of that reasoning is. That diagrams can be produced and reproduced in imagination is an integral part of the creative aspect of abductive reasoning in science, and that reproduction gives rise to a concrete vehicle for communicating the outcomes of such reasoning in intersubjectively testable diagrammatic and logical models.[6]

Acknowledgments Research supported by Estonian Research Council Project PUT267 and the Academy of Finland project no. 12786, *Diagrammatic Mind: Logical and Communicative Aspects of Iconicity*, Principal Investigator Ahti-Veikko Pietarinen. The second author is the main author of Sect. 21.4, the first the main author of other sections. The paper was completed thanks to the 2014–2015 Foreign Experts Program of China at Xiamen University. Early version of the paper, bearing the title "On the Possibility of the Logic of Real Discovery", was presented as a keynote to the *International Symposium of Epistemology Logic and Language*, organized by the Centre for Philosophy of Science at the University of Lisbon, Portugal, in October 2012. We thank the editors, organizers and the audience for comments and discussion.

[6]We leave for further occasion a discussion on the relevance of metaphors in the contexts of discovery, abductive and the model-based reasoning in science, noting only that metaphors are also icons for Peirce: they are the third category of "hypoicons", in addition to the first of images and the second of diagrams (Pietarinen 2008).

References

Baker, V.: Hypotheses and geomorphological reasoning. In: Rhoads, B.L., Thorn, C. E. (eds.) The Scientific Nature of Geomorphology: Proceedings of the 27th Binghamton Symposium in Geomorphology, pp. 57–85. Wiley, Chichester/New York (1996)

Campos, D.G.: Imagination, concentration, and generalization: Peirce on the reasoning abilities of the mathematician. Trans. Charles S Peirce Soc. **45**, 135–156 (2009)

Einstein, A.: On the method of theoretical physics. Philos. Sci. **1**(2), 163–169 (1934)

Feynman, R.P., Leighton, R.B., Sands, M.: The Feynman Lectures on Physics, vol. 2. Addison-Wesley, New York (1964)

Frigg R., Hartmann, S.: Models in science. In: Zalta, E.N. (ed.) Stanford Encyclopedia of Philosophy. http://plato.stanford.edu/archives/fall2012/entries/models-science/ (2012)

Gleick, J.: Genius: The Life and Science of Richard Feynman. Vintage Books, New York (1992)

Hadamard, J.: The Psychology of Invention in the Mathematical Field. Princeton University Press, Princeton (1949)

Hintikka, J.: C. S. Peirce's 'first real discovery' and its contemporary relevance. Monist **63**, 304–315 (1980)

Hoffmann, M.H.G.: 'Theoric transformations' and a new classification of abductive inferences. Trans. Charles S Peirce Soc. **46**, 570–590 (2010)

Hookway, C.: The form of a relation: Peirce and mathematical structuralism. In: Moore, M. (ed.) New Essays on the Mathematical Philosophy of C S. Peirce. Open Court, Chicago (2010)

Kant, I.: Kritik der reinen Vernunft, Riga; Eng. ed. Critique of Pure Reason, translated and edited by P. Guyer and A. W. Wood, Cambridge: Cambridge University Press. Cited as KrV (A/B) (1787)

Nersessian, N.J.: The cognitive basis of model-based reasoning in science. In: Carruthers, P., Stich, S., Siegal, M. (eds.) The Cognitive Basis of Science, pp. 133–153. Cambridge University Press, Cambridge (2002)

Norton, J.: Are thought experiments just what you thought? Can. J. Philos. **26**, 333–366 (1996)

Peirce, C.S.: The Collected Papers of Charles S. Peirce, 8 vols., Hartshorne, C, Weiss, P., Burks, A.W. (eds.). Harvard University Press, Cambridge. Cited as CP followed by volume and paragraph number (1931–1966)

Peirce, C.S.: "Manuscripts" in the Houghton Library of Harvard University, as identified by Richard Robin, "Annotated Catalogue of the Papers of Charles S. Peirce", Amherst: University of Massachusetts Press, 1967, and in "The Peirce Papers: A supplementary catalogue", *Transactions of the C. S. Peirce Society* **7**(1971): 37–57. Cited as MS followed by manuscript number and, when available, page number (1967)

Peirce, C.S.: The New Elements of Mathematics by Charles S. Peirce, 4 vols., Eisele, C (ed.). Mouton, The Hague. Cited as NEM followed by volume and page number (1976)

Peirce, C.S.: Writings of Charles S. Peirce: A Chronological Edition, 7 vols., Moore, E., Kloesel, C.J.W. et al. (eds.). Indiana University Press, Bloomington. Cited as W followed by volume and page number (1982)

Pietarinen, A.-V.: Compositionality, relevance and Peirce's logic of existential graphs. Axiomathes **15**, 513–540 (2005)

Pietarinen, A.-V.: An iconic logic of metaphors. In Proceedings of the 6th International Conference of Cognitive Science, Yonsei University: The Korean Society for Cognitive Science, pp. 317–320 (2008)

Pietarinen, A.-V.: Which philosophy of mathematics is pragmaticism?. In: Moore, M. (ed.) New Essays on Peirce's Mathematical Philosophy, Open Court, Chicago, pp. 59–80 (2010a). http://www.helsinki.fi/∼pietarin/publications/Which%20Philosophy%20of%20Mathematics%20is%20Pragmaticism-Pietarinen.pdf

Pietarinen, A.-V.: Is non-visual diagrammatic logic possible? In: Gerner, A. (ed.) Diagrammatology and Diagram Praxis. College Publications, London (2010b)

Pietarinen, A.-V.: Moving pictures of thought II: Graphs, games, and pragmaticism's proof. Semiotica **186**, 315–331 (2011)
Pietarinen, A.-V.: A realist-modal structuralism. Philosophia Scientiae **18**, 127–138 (2014)
Pietarinen, A.-V.: The science to save us from philosophy of science. Axiomathes **25**, 149–166 (2015)
Pietarinen, A.-V., Bellucci, F.: New light on Peirce's conception of retroduction, deduction, and scientific reasoning. Int. Stud. Philos. Sci. **28**, 353–373 (2014)
Stjernfelt, F.: Natural Propositions: On the Actuality of Peirce's Doctrine of Dicisigns. Docent Press, New York (2014)
Tiles, J.E.: Iconic thought and the scientific imagination. Trans. Charles S Peirce Soc. **24**, 161–178 (1988)
Thagard, P.: How brains make mental models. In: Magnani, L., Nersessian, N.J., Thagard, P. (eds.) Model-Based Reasoning in Science & Technology, pp. 447–461. Springer, Berlin/Heidelberg (2010)

Part VI
Knowledge and Sciences II: The Role of Models and the Use of Fictions

Chapter 22
Does Emergence Also Belong to the Scientific Image? Elements of an Alternative Theoretical Framework Towards an Objective Notion of Emergence

Philippe Huneman

Abstract Emergence is a word that plays a central role in the natural or manifest image of the world, within which we organize our ordinary knowledge. Even though some interpretations of the "scientific image" leave no place for emergence, sciences increasingly made use of this word. But many philosophical arguments have been made against the consistence or validity of this concept. This chapter presents a computational view of emergence, alternative to the usual combinatorial view common among philosophers, that is formulated in terms of parts and wholes. It shows that computational emergence can be characterized in terms of causation, and that a subclass of computationally emergent processes displays many of the connotations of the scientific use of the term. After having so captured a concept of emergence, I turn to the question of applying the concept and testing whether some instantiations exist.

Keywords Emergence • Complexity • Predictability • Causation • Robustness analysis • Computer simulation • Scientific image

It is striking that the theory which holds that only entities of fundamental physics are real entities, and therefore claims that predicates like "think", "believe", "idea", "affection", are illusory – such theory, called "eliminativist" (Churchland 1981), seems immediately absurd to most of us. The question therefore raised by eliminativism is whether the ultimate ontology of everything, given by the sciences, will be at odds with our usual knowledge and way of speaking of things – something like the conflict between two "images of the world" as Sellars put it a long time ago – or whether some of the ontological categories proper to our everyday discourses,

This work has been funded by the Project 'Explabio', ANR # 13-BSH3-0007.

P. Huneman (✉)
Institut d'Histoire et de Philosophie des Sciences et des Techniques, CNRS/Université Paris I Sorbonne, Paris, France
e-mail: philippe.huneman@gmail.com

J. Redmond et al. (eds.), *Epistemology, Knowledge and the Impact of Interaction*, Logic, Epistemology, and the Unity of Science 38, DOI 10.1007/978-3-319-26506-3_22

such as "thoughts", "cars", "trees" etc. have ontological relevance. In this latter case, one of the fundamental insights proper to lay knowledge and everyday discourses is the idea that some "kinds" of stuff are novel regarding some more "basic" things – in the sense there is something novel in a running cheetah that is not included in the quarks that make it up.

The concept of emergence in general aims at making sense of this sense of novelty – of properties, of entities, of laws, etc.[1] – within the framework of naturalism, which seems most accurate to the "scientific image" – namely the refusal of dualism, of positing a region of being besides, and independently of, the natural world as unveiled by natural sciences. The idea of emergence therefore rests on the shared intuition that, if, on the one hand, a scientific mind must not admit any supernatural thing, on the other hand an explanation of things such as trends in economy, thought or affects, or history of political ideas, cannot be worked out in terms of motion of quarks or muons, or other elementary entities in particle physics. For these reasons, the word "emergence" is pervasive in the scientific as well as the philosophical recent literatures. Nevertheless, nothing proves that it would resist a rigorous elucidation; it might be the case that such intuition would fade after an attempt to clarifying it.

In this chapter, after having reviewed some lay uses of the intuitive notion of emergence in the usual discourse and compared it to scientific uses of the term and philosophical traditional reflections on the concept, I will present what I call a computational concept of emergence, contrasting it with another, more frequent, approach to emergence (called here "combinatorial"). I will show that, on the one hand, it is more satisfying and answers better than the combinatorial concept to some objections raised against the very concept of emergence; and on the other side it includes a causal dimension which makes it into a concept proper to capture what is at stake in many appeals to "emergence" in scientific contexts. The last section of the paper will sketch some applications of this concept to special sciences.

22.1 Talking About Emergence: Scientists, Philosophers and Ordinary People

To know things we generally start by partitioning them and the world into various kinds. While these partitions are highly culturally dependent, and vary according to the development of a given individual, it is nonetheless a basic fact of knowledge that we organize all our experiences around a partition into kinds, classify things along these partitions, and acquire knowledge in such a framework. Kant thought that this partition – what he calls the "specification of nature into a logical system"

[1]This is a question left open here – it's enough to point that, following Kim, many philosophical approaches of emergence concern the emergence of properties, even if physicists like Laughlin (2005) talk of the emergence of laws. I argued (Huneman 2008b) that one should first of all speak of emergent processes instead of emergence of properties, these ones being emergent only in a derivative way.

(Kant 1987) – was a basic requisite for any knowledge, since we form "empirical concepts" through comparing and weighting differences and resemblances, and such operation would not be possible if were not presupposing that things will group into kinds, sub-kinds, and so on.

Even if such a partition may vary, in general "living things", "minds", "bodies" constitute an important articulation for them, as well as "animals" and "plants". If there is a natural or "manifest image" of the world, as Sellars (1962) put it, it may clearly include such a subdivision. Cultural anthropology has shown that various cultures will not draw the lines in the same way, many of them having categories in which "life" and "mind" overlap, and a case could be made that it's mostly western thought from the early modern age on that has insisted on a sharp divide between "minds" or "humans" and living beings (Jonas 1966). Developmental psychologists, on the other hand, after Piaget's seminal work (Piaget 1937) accumulated findings about the way western children go through a stage of "animism" where life is a category projected onto all active things, then at age 8–9 restrict this to moving bodies (even falling bodies), and then to bodies that seem endowed with self-motion (e.g. sun, rivers) and finally converge towards an ordinary cultural concept of living things (animal and plants), and then intentional and mental agents.

In addition, on this basis many views have been suggested in order to understand how some entities of a given kind can be articulated with entities of another kind: how human beings can have body and mind, how living things can be generated, etc. For instance animism, vitalism (Wolfe and Normandin 2013), and mechanism are families of theories that articulate differently an understanding of what life and livings things are, and how they connect to physical things. Concerning mind and mental states, philosophers have been designing varieties of monism, dualism, panpsychism – even though in many non-Western cultures we fail to see the way we westerners take for granted that dualism has to be taken into account (Descola 2005), so that the "body and mind" problem so familiar to contemporary philosophers of mind does not make sense.

Therefore, if we want to roughly sketch the framework for a natural image of the world that is more or less shared by many cultures, and in which people organize their knowledge of things, there is an important room for a notion of how things of one kind may arise on the basis of other extant things. "Emergence", defined as the "progress of coming into existence or prominence" by the Oxford dictionary, plays this role in scientific and philosophical contexts. It is interesting to see that etymologically it derives from Latin word "*emergentia*" which means "coming to light": the ordinary concept, then, carries this connotation that what is emerging was somehow concealed in what it emerges from, or in other word, that what emerges comes from something that had a potential for making it emerging. Many theories that have been elaborated in the past concerning the existence of organisms on the basis of brute matter – "spontaneous generation" – assume that life *emerges* from dead or brute matter.[2]

[2]See Roe 1981 on the entanglement of spontaneous generation idea with controversies over generation.

Thus ordinary ideas of the nature of main kinds of things in the "natural image" of the world in which our usual knowledge develops include a room for this intuitive idea of emergence. However, a sense of emergence is not at all absent from the modern scientific image: emergence talk indeed occurs precisely when it comes to account for the fact of novelty, or novel kinds of things in a given field. Scientists often use the term "emergence" because many of them believe that even if the investigated phenomena are made of material items which obey laws of elementary physics, such physics is not sufficient to understand them. This is true for human and social sciences but also for biology and even for the regions of physics which are not particle physics (the so-called "fundamental physics"). For philosophers on the other hand, emergentism first means a view defended in the 1920s by Samuel Alexander, Lloyd Morgan or C.D. Broad, philosophers who thought that emergence actually reconciled naturalism with the acknowledgment of the existence of novel properties beyond elementary physics. They were proved wrong by the progress of science to the extent that one of their paradigmatic examples was the properties of water – which were, according to them, unexplainable through atomism – and which later have been explained precisely by the quantum physics of covalent bonding (Mc Laughlin 1992). The notion then came back through the field of the philosophy of mind: the main problem here is to understand mental states as, at the same time, grounded upon, and irreducible to, brain states. The discussion then revolved around argument suggested by Jaegwon Kim, who sees mental properties as epiphenomenal ones, because if one is committed to the "causal closure of physics",[3] they cannot have any causal efficiency (all their causal strength comes from their physical bases) and then they have a mere epiphenomenal reality. However, the generality of the use of the word emergence in the sciences (E.g. Anderson 1972; Laughlin et al. 2005) contrasts with the specificity of the use of this term by many philosophers. Some of them, considering the concept, come to the conclusion that either there are no emergent properties at all, or only phenomenal consciousness (i.e., "what it's like" to have this thought or to be this person, e.g. Nagel (1974): "what it's like to be a bat") would be emergent (Chalmers 2006). But, following the general orientation of the volume edited recently by Bedau and Humphreys (2008) I aim here at making sense of the concept of emergence as one can find it in the sciences, instead of discussing what should be the concept of emergence and what would instantiate it in the sole light of the mind/body problem.

As said Bedau (2008) a concept of emergence must at the same time mean the *autonomy* (viz. some bases) and *dependency* (upon these bases) of what is emergent. An important distinction has to be made between two different questions regarding emergence, namely the question about the *meaning* of emergence, and the issue of the *reality* of emergence. The former is about building a coherent concept of emergence, likely to capture many of the uses of the word in the sciences. The latter is whether there are things in the world that actually fall under this concept.

[3]Idea that any physical fact or event has a cause which is also physical – notwithstanding what other facts or causes may exist. This postulate is supposed to be inherent to modern science.

This distinction is necessary, because many arguments in philosophy – first of all by Kim – were directed against the consistence of the concept of emergence, i.e. showing that either it makes no sense or it means some kind of epiphenomenalism. In this perspective, if Kim and his supporters are right what we call "emergent" is not emergent because the very concept of emergence is misconstrued, and therefore the question of checking whether something in the world falls under this concept makes no sense.

On the contrary, it is conceivable that we could devise a satisfying concept of emergence, and that in the end nothing empirically falls under this concept – even though in some other possible worlds some possible things may fall under the concept. Construing the concept is the first philosophical question; testing whether what is believed to be emergent and then falling under such concept, actually falls under this concept, and finally whether there exists in the world something which belongs to the extension of such concept, is another question, to be mostly answered by the empirical sciences. Some confusion occurred in the debates because these two questions, relevant to two kinds of investigations, have been conflated. Thus, I mostly here elaborate a concept of emergence, and only the last part of the chapter deals with whether or not something falls under this concept and how we can know it. I start by elaborating a concept of emergence that seems to me valid, and also that is such that satisfying this concept proves to be objective or independent from our cognitive abilities. Then it is shown that, by definition, what satisfies this concept is unpredictable, and then I show that a specification of this concept captures what seems emergent to us in many uses of scientific talk.

22.2 Combinatorial and Computational Emergence

22.2.1 Characters of Emergence and the Non-triviality Requisite

Emergence is often conceived of as the issue of understanding the properties of a whole which would be irreducible to properties of the parts – what I call "combinatorial emergence". Traffic jams (Nagel and Rasmussen 1994), fads (Tassier 2004), temperature, chromosomes at the time of meiosis, all display a behavior which cannot be understood by adding accounts of the behaviors of their parts. It entails that one sees them as "emergent behavior": such emergence is often viewed as something proper to the whole and irreducible to the parts. Philosophers like Silberstein (2002), O'Connor (1994), Bechtel and Richardson (1992) tackled the issue of emergence through this scheme of the parts vs. whole. In the same way, Phan and Dessalles (2005) see emergence as a drop in complexity, J. Wilson (2010)

as a decrease in degrees of freedom – contrary to the mere product of properties of parts (where there would be additivity of degrees of complexity/degrees of freedom[4]).

But let's take what is often seen as a famous example of emergence, the segregation model by economist Thomas Schelling (1969): according to their "colour", agents in an agent-based model will eventually get lumped together in homogeneous clusters, like ghettos in real life. The rules are, as one knows, only to have a slight dislike for being part of the minority (something like "if I am the only green among ten reds, then move"). Besides the important teachings of this model in social sciences (essentially about the limits of a desegregation politics based on education), the behavior "join the group" is not given in the behavioral rules of agents, so it is somehow emerging from the added interactions. But groups are not exactly *composed* of agents, because those groups subsist even if some agents are added and some emigrate or die (Gilbert 2002). Therefore, given that parts are transient regarding the whole, a simple view of emergence as irreducibility of the whole to its parts is mistaken.

Philosopher William Wimsatt (1997) defined emergence as the "failure of agregativity". The main issue here is to provide then criteria of agregativity – which means tackling the issue of emergence in an inverse way. Wimsatt's criteria for failure of agregativity are a sophisticated formulation of what is happening when we say that we cannot reduce the properties of the parts to those of the whole. These criteria are: invariance through substitution of parts; qualitative similarity through addition or subtraction of parts; invariance regarding decomposition-re-aggregation of parts; lack of cooperative/inhibitory interactions. Those are criteria of invariance; thereby they take into account the case of parts which change and alternate in a whole like in the segregation model. However, it seems that, except mass, almost nothing is genuinely aggregative, namely satisfying all these invariance criteria. This is clearly a problem for the combinatorial view. Emergence should surely be ascribed to fewer properties than "everything, except the mass"; therefore it should require an additional criterion which is not provided by such analysis. Here we are left with the idea that emergence comes by degrees.[5] However in such view, the meaning of emergence is quite superfluous, it would be enough to talk of degrees of agregativity; the concept of emergence can only have a meaning with this additional criterion according to which emergence is more than a mere lack of agregativity, but precisely it can't be provided by the combinatorial view.

Actually, emergence is supposed to encompass several characteristics: unpredictability, novelty, irreducibility (Klee (1984), Silberstein (2002), O'Connor (1994), Crane (2001), Chalmers (2006), Seager (2005), Humphreys (1997) largely concur on these characteristics). Many add "downward causation", but this is more controversial. Irreducibility understood as irreducibility of properties of the whole to properties of the parts seems now quite trivial given the previous considerations,

[4] See Atay and Jost 2004, 18.

[5] Also Bechtel and Richardson 1992.

and too frequent to provide something as "emergence".[6] Concerning novelty, since the properties of the whole are quite always novel regarding properties of the parts – think of colour, or even volume ... – the real issue is: *which* novelty should count as emergent? We are left once again with no objective criterion. "Novel" most of the times mean what has no name yet in our language (Epstein 1999). Hence this unavoidable characteristics leads to a widely shared conclusion: if emergence has any meaning, it is restricted to epistemological emergence, namely relative to a set of theories and cognitive abilities – perhaps to the exclusion of the exceptional case of *qualia* (Chalmers 2006; Crane 2001; Seager 2005; O'Connor 1994). Generally, most of the authors oppose epistemological and ontological emergence (the latter being in the real world, the former being defined by the weakness of our analytical or theorizing abilities). Most would conclude that the concept of emergence is undoubtedly epistemological only. A major argument for this conclusion is that, as the example of water for British emergentists can remind it, that what seems now emergent is such only relatively to our theories, and that nothing precludes that a more sophisticated theory could later explain how – to stay in the framework of combinatorial emergence – the properties of the whole result from properties of the parts, or are simply the conditional properties of parts, now actualized. Another argument is the fact that what is real must have causal properties, yet if emergent properties emerge upon some bases, and are not transcendent, they receive their causal powers from those of their bases, so they don't have any of such powers on their own, and thus don't have a proper ontological character. Kim's arguments of exclusion and overdetermination provide the most achieved form of this argument.

The rest of this chapter explores another concept of emergence than the combinatorial one; I show that it is immune to the triviality problem revealed by Wimsatt's non-agregativity criteria, and to the usual verdict that emergence is eventually epistemological, and emergent properties are epiphenomenal.

22.2.2 The Incompressibility Criterion and Emergent Processes

In the framework of computer simulations one has defined what Bedau (1997) calls "weak emergence".[7] According to the purported criterion, a state in a computational process is weakly emergent if there is no shorthand to get to it, except by running the simulation. ("The *incompressibility* criterion of emergence" – see Huneman 2008a, b; Humphreys 2008; Bedau 2008; Hovda 2008). This approach, amongst the four mentioned connotations of emergence (unpredictability, irreducibility, novelty, downward causation), starts from the notion of unpredictability.

[6] See also Bar Yam (2004).

[7] Humphreys (1997) is the first systematic investigations of epistemological problems raised by the generalized use of simulations in the science. Huneman (2011; 2014) tackled this problem in the framework of evolutionary explanations.

Such an approach bypasses the question of the cognitive subjectivity proper to the novelty problem in the former approach, because it's based on a computational property of algorithmic models. That is why we would have a major clue about emergence which would be, if not ontological, at least objective in the same way as conceptual truths of mathematics are objective, i.e. independent of our cognitive capacities or epistemic choices.

Yet, one could object that our criterion of incompressibility is only *temporary*, because we cannot claim that in a remote future, with increased computational capacities, we will be still unable to find analytical shortcuts to reach faster the final state than by simulation. However, here is the sketch of a refutation of such objection. Huneman (2008a) develops some arguments in favor of the objectivity of computational criteria on the basis of Buss et al. (1992). The basic idea consists in building a set of logical automata whose values change according to a global rule ***R***. Each automaton transforms the value of its cells according to an input 0 or 1. Applying the global rule ***R*** depends upon the numbers of each values (q_1, q_2 . . .) in the set of automata at step n; the input function which determines then the input of all automata at step $n + 1$ is determined by ***R.*** For this reason the system is perfectly deterministic.

For a class of rules, it can be shown that the problem of predicting the state of the automata set at time T arbitrary remote is PSPACE complete (see Box 22.1). This result perfectly illustrates the fact that some computational devices are objectively incompressible. As authors write: "If the prediction problem is PSPACE complete, this would mean essentially that the system is not easily predictable, and that there seems to be no better prediction method other than simulation" (Buss et al. 1992, 526) Even with infinite cognitive capacities, there would be a real difference between predictions problems which are PSPACE complete and others, therefore the computational definition of emergence is *objective.* Weak emergence so defined as inaccessibility except via simulation is then not something trivial since, in this context, all global rules which are constant-free are computable in polynomial time, which makes a clear distinction between weakly emergent cases and other ones.

Box 22.1: Complexity Classes of Prediction Problems for Automata
Input function:

If $Z_n = 0$, $F(n + 1) = g_0(F(n))$
If $Z_n = 1$, $F(n + 1) = g_1(F(n))$

Functions g_0 and g_1 have their values in $\{q_1 \ldots\ldots q_j \ldots\ldots q_n\}$.
Global rule ***R***: Z_i has it s values in {0,1}.

(continued)

$Z_i = \text{Max}\ (N_i\ (q_1) \ldots\ldots\ldots N_i\ (q_j) \ldots\ldots\ldots N_i\ (q_n))$ where $N_i\ (q_j)$ is the number of times the value q_j is taken at step i.

Step 0	F_1 (0)	F_2 (0)	F_i (0)	F_m (0)
Step 1	F_1 (1)	F_2 (1)	F_i (1)	F_m (1)
Step k	F_1 (k)	F_2 (k)	F_i (k)	F_m (k)
.				
.				
Step n	F_1 (n)	F_2 (n)	F_i (n)	F_m (n)

Some global rules are constant-free, meaning that they can be formulated with no reference to one of the real values $q_{i\ldots..}$of the constants, and the other cannot."If there are as many q_i as q_j, and for all values of i and j, let $Z = 0$; otherwise $Z = 1$" is an example of a constant-free rule. Buss et al. (1992) have shown that, if the global rule is constant-free, then the problem of predicting the state of the system at time T is PSPACE-complete; that is why the problem cannot be solved in polynomial time (since we assume that no $P = NP$ and that NP problems are included in PSPACE problems, so that being PSPACE complete implies being a problem such that all other problems can be solved if such problem can be solved, which makes such a problem at least harder than NP-complete). A detailed demonstration rests on the fact that constant-free global rules are preserved for any permutation of q_i....., which constitutes a major difference concerning the computational pattern of prediction.

22.3 Causation and Computational Emergence

The present approach starts with a concept of emergence to show its coherence and plausibility. Another issue is then to decide whether exist some things which, in the real world, fall under this concept, that is, are such that if one has an accurate model of the phenomenon, the model will display properties of computational emergence. It's conceivable, for now, that there are none, or that we don't know whether the current models we have, and which speak for the existence of emergent properties, are accurate enough. What is shown until now, is that with incompressibility one has a non-trivial, objective, concept of emergence. Because I am only concerned here with the meaning of emergence and not its actuality, I rely on formal properties of simulations such as cellular automata or genetic algorithms. We can't rely solely on them to find out instances of the concept of emergence in the world, but here they can allow us to construe and justify a proper concept of emergence.

Such concept, starting from the idea of unpredictability, includes the notion of irreducibility. I will now show that the notion of novel order included in the intuitive

notion of emergence can be made sense of through this computational concept. To this end, I show first that one finds in the computational concept of emergence a dimension of causation, so that it's not a mere formal notion, whatever the degree to which this concept is instantiated in the real world. From this on, I show (2.3) that the connotation of novel order is likely to be met for a subclass of systems displaying computational emergence. (This section surveys arguments presented in Huneman 2008b). The last section (3) will give some clues for the issue of finding in the real world instantiations of this concept of computational emergence.

22.3.1 Causation and Simulations

First, this is about answering the objection that starting from the context of simulations to conceive of emergence compels one to leave aside the most important thing, namely the fact that one calls "emergent" real processes, which thereby encompass some causation, and possibly raise the issue of *downward causation*, that is, emergent properties of entities (e.g. mental states, fads, standing ovations . . .) causing backwards effects in their bases (brain states, agents, individual spectators). Peter Corning enunciates such objection very clearly, criticizing in general approaches of emergence relying on computer sciences, such as Holland's views (*Emergence* 1998), based on a study of genetic algorithms: "Consider Holland's chess analogy. Rules or laws have no causal efficacy; they do not in fact "generate anything". They serve merely to describe regularities and consistent relationships in nature. These patterns may be very illuminating and important, but the underlying *causal agencies* must be separately specified (though often they are not)." (Corning 2002, 26).

Yet, simulations actually can include a dimension of causation. Let's take first cellular automata. A cellular automaton is a set of cells which can be in several possible states, the state of each cell C at time $n + 1$ being determined by its state at n, through a rule which assigns a state to C at $n + 1$ according to the state of neighboring cells of C at n. This system is wholly deterministic.

Actually, I argue that there are relations of causation within simulations, which are given by the specifications of properties at successive times in the simulation: some properties of a cellular automaton at time $n + 1$ have on the background of the rules a sole cause, namely the properties of it at time n. This argument uses the counterfactualist concept of causation, first elaborated by Lewis (1973) and refined since (e.g. Hall and Paul 2004). According to this concept, A causes B iff "if there had not been A, there would not be B." (I leave aside some subtle distinctions, which aim at excluding obvious counterexamples from this rough formulation of causation.)

Since the rules of the cellular automaton are such that several neighborhoods of the cell C (i; n) yield the same state for C (i; $n + 1$), one can't say that "if there had not been the same neighborhood state C ($i - j$, $i + j$; n) there would not be C (i, $n + 1$)". However there are properties which give rise to a counterfactual dependence between their instantiations at n and at $n + 1$, as sketched in Box 22.2.

Therefore characterizing a class of simulations as computationally emergent should entail a specification of this class in terms of a specific causal pattern.

Box 22.2

Let's call $A_1{}^n$ a set of states of cell $a_{1+m}{}^n$, m varying from 0 to p (p defined by the rules of the CA), such that their result at level $n+1$ is the state (in the considered CA) of $a_1{}^{n+1}$. But there are j other sets of states like $A_1{}^n$, such that their outcome is always $a_1{}^{n+1}$. We can write the set of these sets of states $A_1{}^{n,k}$, k varying from 1 to j. The property $P^{n,1}$ is then defined as such: a CA is said to have property $P^{n,1}$ at step n iff it exists k, $0<k<j+1$, such that it is in a state belonging to a set of states $A_1{}^{n,k}$.

Now, for any i, and for a given state of $a_i{}^{n+1}$ at step $n+1$, one can define a set of states $A_i{}^{n,j}$, all of which result into the considered state of $a_i{}^{n+1}$, and then define the property $P^{n,i}$. let's define Q the property of being in the state $\{a_1{}^{n+1} \ldots a_i{}^{n+1} \ldots a_m{}^{n+1}\}$, property instantiated by the CA at step $n+1$. Finally we can write that the CA is P at step n iff it has all the $\{P^{n,i}\}$. Then, it is true that: "if the CA had not had property P at n it would have not been Q at $n+1$" – this is a counterfactual dependence, hence a causal relationship. So, the causal explanation of "having property Q at step $n+1$" is "having P at n". Thereby there are, in simulations such as CA, causal relations between sets of states at different steps.

Causation in CA as counterfactual dependence between properties at some steps. (After Huneman (2008b)*).*

22.3.2 Causation and Incompressibility. Emergence as Break Up in Causal Explanation

Those counterfactual correlations belong to the set of all cellular automata. But when there are emergent properties, this means a specific causal singularity. Quickly said,[8] when there is emergence, it means that the causal relationship between two successive states of the system ($A^i{}_n$ and $A^i{}_{n+1}$ for all i) can never be traced back to a global law of the system (for example a law for all A_n). For a cellular automaton, one can go from $a^i{}_n$ (with j varying from $i+p$ to $i-p$, p being defined by the rules of the CA) to $a^i{}_{n+1}$, but not in general in a nomothetic way from A_n (the set of states $a^i{}_n$) to $A_{n+1,}$ and even less from any given n to A_n. This could be a way to make sense, from the viewpoint of computational emergence, of what Wimsatt called failure of

[8] Demonstration in Huneman 2008b.

agregativity, because any aggregative system is such that we can go in a somewhat continuous way from the local to the global, and write a global law to describe this process.

In general, for any system the causal explanations can be of two fashions – either forward on the basis of the elements (forward-local), or backward from the whole (global backward). About a thrown stone, one could write the position of the ball at any instant by deriving it from a trajectory given by the law of gravitation, or compute the position, instant by instant, according to its prior position. This problem in dynamics is such that both approaches coincide because the trajectory is often integrable. The coincidence between causal explanation means that the step by step explanation, represented by a running cellular automaton, and the explanation by a rule, which represents the jump from the initial state to a step n of the automaton and is given by the motion's equations, do coincide. When there is computational emergence, we lack such a coincidence: in this sense, the proper character of causation in simulations that represent emergent (in the computational sense) processes, is indeed this fracture within causal explanation. If actually such a coincidence was always by principle available, then we would always have a rule to go from the local (explaining the a^i_{n+1} by the $(i-k<j<i+k)$ a^j_n) to the global, whereas, as we just saw it, this is not the case for emergence because incompressibility means the lack of a shortcut that would play the role of the global equation allowing to fit the global-backward and the local-forward explanations. The *causal signature* of computational emergence is thereby the break up between those two modalities of causal explanation.

22.3.3 Emergent Causal Reliability and Emergent Order

Nevertheless there is, *among* computationally emergent phenomena, a *subclass* of processes such that, beyond some step, regularities between sets of cells arise. For example, think of gliders and glider guns in Conway's Game of Life, which is a two-dimensions cellular automaton whose rules are given in Box 22.3 (Gardner 1970).

Box 22.3: Rules for the Game of Life

Consider a two-dimensional cellular automaton, in which each step can be in two states, *alive* or *dead*. The transition rules from step n to step $n+1$ are:

At each step in time, the following transitions occur:

1. Any live cell with fewer than two live neighbours dies, as if caused by under-population.
2. Any live cell with two or three live neighbours lives on to the next generation.

(continued)

3. Any live cell with more than three live neighbours dies, as if by overcrowding.
4. Any dead cell with exactly three live neighbours becomes a live cell, as if by reproduction.

Gliders and glider guns are common terms for recurring patterns that occur within it, and that look like flying gliders thrown from a stable device (Fig. 22.1) Here, we could say it in a counterfactual way: if the set of states (defining a glider or glider gun) had not been there (in this position), the glider (as a set of cell-states) would not be in the state one finds it.

These dependencies between partly global states of the simulation are not given with the initial rules, which concern only sets of individual cells. In the usual sense, these emerge in the course of the simulation, and can concern a mere transient state of it. But when they happen, they allow a much more simple explanation of the behavior of the simulation than appealing to the rules that govern each cell's behavior. Why simple ? Because usually one has to specify the states of all cells in order to step by step explain the simulation, whereas here when a rule (as a transient counterfactual dependency) has emerged, it can be stated by specifying positions of sets of states only. This is a coarse-grained explanation (gliders flying, loop self-replicating in Langton's loop improved by Sayama[9]), which of course

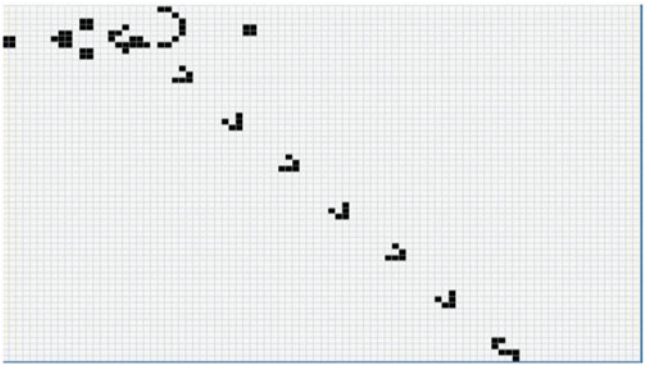


Fig. 22.1 Gliders in a Game of Life simulation. In this grid, cells are either *white* or *black*, and the state of *a* is determined by the state of the parent cells (*white/black*) and its eight neighbors according to a rule. Gliders are these patterns of *black dots* extended through several lines that are conserved as such along many steps of the simulation, therefore that seem to "move" (translate while rotating) regularly through the grid towards the *bottom right*, even though the cellular automaton only determines the state of cells at each step of the run of the simulation

[9]Langton 1989. On this loop see Salzberg et al. 2003; Sayama 1998.

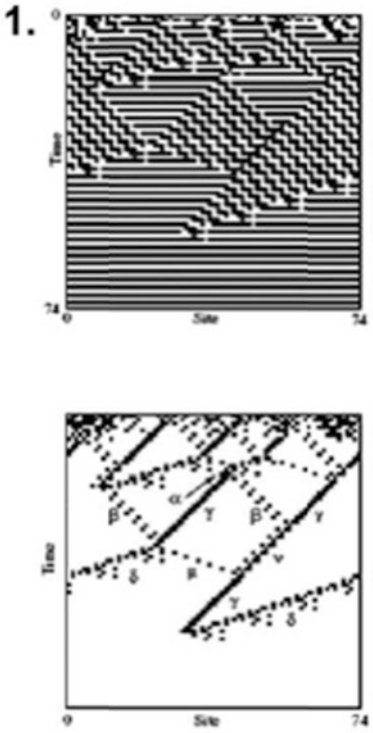


Fig. 22.2 Filter (Crutchfield and Hanson 1993) intending to reveal "mechanical" entities (greek letters in *bottom* diagram) which causally explain it

omits details, but in the same time saves both computation time and information, and allows reliable generalizations, like Epstein's civil violence study (see below). Israeli and Goldenfeld (2004) have shown that most of the CA rules support, at some point, to be formulated as coarser grain rules, the sets of cells being then taken as cells, so that apparently incompressible rules can be translated, through a coarse grained description, into compressible rules. (Of course in many cases the coarse grained cell *is* what indeed emerges, *sensu* the incompressibility criterion, in the simulation).

More technically, Hanson and Crutchfield (1997), Crutchfield and Hanson (1997), Crutchfield and Shalizi (2001) developed a method for reading CA in terms of "mechanics": they identify patterns (named, by analogy with mechanics: lines, points, particles, etc.) whose correlations are such that they underwrite the running of the whole CA (Fig. 22.2) and can even be automatically detected (Shalizi et al. 2006). Here, clearly we get a causal lexicon that enables one to make sense of the intuition that emergent properties are such that they encompass another kind of causation. This novel causality is, first of all, counterfactual regularity between sets of cells (in CAs) or agents (in ABMs[10]). In such a case, reliable predictions and

[10]For emergence in ABM according to my criteria, see R. Wilson 2010.

descriptions of the system can be given at the level of this novel causality, using such dependencies (hence the term "reliable causal emergence" I used to name this subclass of computational emergent processes).

22.4 Applications

This *reliable emergence* – in the sense of the subclass of emergent processes satisfying the clause 3.3. of causal reliability – is instantiated by numerous models in empirical sciences. Take the study of fads (Tassier 2004). In this case, one can see clear relations of causality between states of fads at some moments, which define a general pattern of fads processes. According to the parameters, in these multi-agent models the agents either separate into clusters, any of which adopting a given fashion – or display a behavior such that cycles of fads become visible.

In the same way, emergence of local norms (Burke et al. 2006) appears as a computationally incompressible process leading to specific patterns, possibly alternate, or fix. The whole system satisfies the computational emergence criterion because one can't analytically derive the result. Also, traffic jams (Nagel and Rasmussen 1994) display patterns whose arising process, modeled by a CA, is incompressible. In the same way, once a traffic jam has appeared, it is likely to show causal relationships between itself and, either other traffic jams, or some state variables of the system such as the average speed of cars etc. Finally lipid membranes (Rasmussen et al. 2002) satisfy the same criteria concerning their creation, the authors describing as a discrepancy between two languages the same difference here described as a distinction between a global rule and the lack of immediate transduction into global rules in the case of emergent processes.

With Epstein's work on civil violence (Epstein 2002) one sees a peculiar kind of counterfactual dependence.[11] In these models, Epstein defines agents representing social individuals, and studies their propensity to rebel. Each model implements quite intuitive rules, according to which the acting out (violently) of an agent against the State depends both upon the perceived risk, and the frequency of agents already acting out in her neighborhood. This is a typical multi-agent model. Two parameters are defined: level of oppression and level of legitimacy of the government. By varying these parameters, and multiplying simulations, Epstein can show numerous counterfactual dependencies between the values of parameters (or their variations) and global outcomes of simulations (namely the frequency and generalization (or not) of a rebelling behavior.) Here, dependency does not take place between two moments of the simulation, but – on the basis of a large set of simulations – between range of values of parameters and classes of global outcomes (see also Huneman 2012). This latter dependency is a generalization of the former one, hence causation

[11] A more precise description of levels of counterfactual dependency, defining modes of regularity and prediction, is done in Huneman (2012).

(counterfactual) at an upper level. One of the most striking results of this study is that, when the legitimacy of a government drops, this increases the probability of a violent uprising – however the relevant variable here is not the width of the drop, but its speed: a small but quick drop of legitimacy more easily entails an uprising than a much larger but slower drop.

22.4.1 Robustness Analysis

A last concern could be the following. I have shown that computational emergence is objective, that it concerns some specific causal explanations, that it allows one to define a subclass of emergent phenomena (called 'reliable emergence' here) displaying causal relations such that one can recover the idea – proper to our intuitive conception of emergence – of a novel and spontaneous order. Yet someone could always make the following objection: even if such a notion is formally correct, when one asks what instantiates it in the real world, some phenomena such as the ones described indeed instantiate it *only under the condition* of our accepting the models which describe them. In other words, if this concept of emergence is ontological, it is however not buffered against the fact that nothing would instantiate it in the real world because all models which make us conclude that it is instantiated can, one day, be superseded by models with no emergence. Many arguments have been used against such objection (e.g. Bedau 2008; Humphreys 1997, etc.). I indicated (Huneman 2012) the idea that robustness analysis as done by scientists usually would provide a way to conclude positively regarding the genuinely emerging character of some phenomena, independently of the model.

The idea of robustness analysis is the following (Levins 1966; Weisberg 2006). To build a model implies choosing parameters and ascribe values to them; this choice of course often involves simplifying assumptions, namely, betting that some parameters are not so relevant to the phenomenon. Behaviors and general outcomes of the model can vary with those values, and more generally with the choice of parameters themselves. In Epstein's model of civil uprising for example, education is not a parameter. One could add it, and then check whether the model behaves similarly when education level varies. We say that an agent-based model is robust if its range of qualitative behaviors does not change when the parameters themselves change. For example if education would not change anything to Epstein's results, this would speak for the robustness of the model (across known parameters).

Levins (1966) original paper contrasts a model of the evolution of some trait in a population approached via an analysis of gene frequencies, and another model approached as an optimizing process of the phenotypic values. In his perspective, the former (A) is more realist (since it takes into account the genetic make up of organisms) whereas the latter (B) is more general (since it could be applied to various species that have these traits). His general claim is that models indeed may

privilege either realism, or generality – or even precision, e.g. predictive accuracy – one over the other since all these epistemic values can't be maximized at the same time. It results that the "theorems" that can be derived in both models, e.g. (A) and (B), are called "robust" theorems, and Levins contends that they have good chances to be verified in the real system under study. Actually both Levins' claims have been challenged: some authors argued that the idea of being compelled to trade-off epistemic values is overstated (Orzack and Sober 1993), while others argued that "robust theorems" are not, as such, likely to be "true" but rather, they need additional empirical confirmation to be taken as true (Kuorikoski et al. 2012). Yet, this view triggered many reflections, beyond the epistemology of ecology, especially about the kinds of trade-off between epistemic values that characterize different fields or subfields (e.g. Matthewson and Weisberg 2009, on economy).

Let's go back to our examples of agent-based models, and assume that we have a model of a target system with a given set of parameters, which satisfies the two emergence clauses and the reliability clause. And let's now assume that testing most of the parameters, adding some and changing some of them, there is still the finding that reliable computational emergence occurs in models for some values of the variables. This finding is therefore robust across models and, according to Levins' general argument, that means that it can somehow corresponds to reality. The phrase "being true", or "capturing reality", is as such not very precise, and for the present purpose I take it that it means, among other notions, the idea of "capturing the causal structure of the phenomenon", as Wimsatt (1987) emphasized. Many arguments then support the idea that such propositions robustly established through a family of models describe the causal structure of the phenomenon: one can appeal to an argument of the type "inference to best explanation" (Lipton 1991) in order to reach such conclusion (i.e. nothing explains the fact that a same proposition is derived across many models embedding a variety of parameters better than the fact that the content of such proposition is true), or one can conceive of an argument stating that causation is based on counterfactual dependencies (Lewis 1973; Hall and Paul 2004), and, since the model displays dependencies between intervening variables and then causality (Woodward 2003), robust findings in the model capture these dependencies.

From this on, such set of agent-based models, describing the causal structure of reality through robust findings, and satisfying criteria for emergence, makes it legitimate to say that the process at stake in the system under study presents emergent properties. More precisely, since the signature of emergence is a certain specificity of the causal structure – as seen in the preceding section – it is clear that if a finding in a model captures the causal structure of reality and displays emergence, then the causal structure of reality will carry the signature of emergence and therefore real processes will be emergent (be it reliable emergence or mere emergence). In other words robustness analysis of the models is the last step that allows one to infer in some given case, from the satisfaction of formal emergence criteria to the reality of emergent processes or properties.

22.5 Conclusion

This chapter explored various aspects of a theory of emergence based on the idea of computational emergence. It starts from the everyday use of the word emergence, which has an important role in the natural image in which we construct our knowledge. It is then argued that in the sciences the concept of emergence is increasingly used to make sense of novelty, which is pervasive in the natural image, and then the question arises of whether this is a consistent and well-formed concept. The philosophical analysis posed here intends to answer this question by showing that there is a rigorous way of forging such a concept of emergence that is likely to do the job it is expected to do in various scientific contexts. To this extent, the sciences do not in principle contradict the daily knowledge (embedded in the natural language) that some things/properties emerge. Emergence may not be the place where the scientific image breaks up with the natural image of the world. However it was shown that the concept of emergence adequate to scientific practice has to be construed in a very specific manner: computational incompressibility is the proper criterion for emergence, because the combinatorial approach suffers from the triviality problem. Especially, the concept of emergence so construed is objective and non-epistemic, in the sense that it does not depend upon cognitive capacities and achievements of the subjects. Nevertheless, such concept is not restricted to characterizing purely formal relations, since it can specify the *causal* signature proper to emergence as a breaking between both aspects (forward-local, backward-global) of causal explanation.

In this perspective, what corresponds to the intuitive notion of emergence found under various guises in the natural image is the subclass of processes that are computationally emergent and that, moreover, were characterized above as reliable causal emergence (i.e., that feature at some point coarse-grained counterfactual dependencies). At least the world of numerical simulations provide numerous examples of that, from the Game of Life trough Langton's loop to Holland's Echo (Holland 1995, 1998).

Philosophy can't say much more about emergence: whether this concept is instantiated or not in the real world is an empirical question, to which only science can answer – and, the answer seems to be affirmative, regarding the examples given here. But considering scientific models from the computational perspective allowed us at least to bypass metaphysical objections against the concept of emergence itself. The last philosophical point that can be made concerns an answer to the skeptics' objection contending that anything assertable about our models is subject to a principled suspicion when applied to the reality that they model. In effect, robustness analysis combined with arguments of the kind "inference to best explanation" is likely to make the computational concept of emergence into a concept that is applicable to state of affairs in the world. From there, the question of what instantiates such concept, and how we can prove that it is indeed instantiated in the world, is just beginning.

References

Anderson, P.W.: More is different: broken symmetry and the nature of the hierarchical structure of science. Science **177**, 393–396 (1972)

Atay, F., Jost, J.: On the emergence of complex systems on the basis of the coordination of complex behaviors of their elements. Complexity **10**(1), 17–22 (2004)

Bar Yam, Y.: A mathematical theory of strong emergence using multiscale variety. Complexity **9**(6), 15–24 (2004)

Bechtel, W., Richardson, R.: Emergent phenomena and complex systems. In: Beckermann, A., Flohr, H., Kim, J. (eds.) Emergence or Reduction? pp. 257–287. de Gruyter, Berlin (1992)

Bedau, M.: Weak emergence. In: Tomberlin, J. (ed.) Philosophical Perspectives: Mind, Causation, and World, vol. 11, pp. 375–399. Blackwell Publishers, Oxford (1997)

Bedau, M.: Is weak emergence just in the mind? Mind. Mach. **18**, 443–459 (2008)

Bedau, M., Humphreys, P.: Emergence. Contemporary Readings in Philosophy and Science. MIT Press, Cambridge (2008)

Burke, M., Furnier, G., Prasad, K.: The emergence of local norms in networks. Complexity **11**(5), 65–83 (2006)

Buss, S., Papadimitriou, C., Tsisiklis, J.: On the predictability of coupled automata: an allegory about chaos. Complex Syst. **5**, 525–539 (1992)

Chalmers, D.: Strong and weak emergence. In: Clayton, P., Davies, P. (eds.) The Re-emergence of Emergence, pp. 244–256. Oxford University Press, Oxford (2006)

Churchland, P.M.: Eliminative materialism and the propositional attitudes. J. Philos. **78**, 67–90 (1981)

Corning, P.: The re-emergence of "emergence": a venerable concept in search of a theory. Complexity **7**(6), 18–30 (2002)

Crane, T.: The significance of emergence. In: Gillett, C., Loewer, B. (eds.) Physicalism and Its Discontents, pp. 207–224. Cambridge University Press, Cambridge (2001)

Crutchfield, J., Hanson, J.: Turbulent pattern bases for cellular automata. Phys. D **69**, 279–301 (1993)

Crutchfield, J., Hanson, J.: Computational mechanics of cellular automata: an example. Phys. D **103**, 169–189 (1997)

Crutchfield, J., Shalizi, C.: Pattern Discovery and Computational Mechanics. arXiv:cs/0001027v1 (2001)

Descola, P.: Par-delà nature et culture. Gallimard, Paris (2005)

Dessalles, J.L., Phan, D.: Emergence in multi-agent systems: cognitive hierarchy, detection, and complexity reduction. In: Mathieu, P., Beaufils, B., Brandouy, O. (eds.) Artificial economics. Lecture notes in economics and mathematical systems, vol. 564, pp. 147–159. Springer, Berlin/New York (2005)

Epstein, J.: Modeling civil violence. Proc. Natl. Acad. Sci. U. S. A. **99**(3), 7243–7250 (2002)

Epstein, J.: Agent-based computational models and generative social science. In: Generative Social Science: Studies in Agent-Based Computational Modeling, pp. 4–46. Princeton University Press, Princeton ([1999] 2007)

Gardner, M.: The fantastic combinations of John Conway's new solitaire game "life". Sci Am **223**, 120–123 (1970)

Gilbert, N.: Varieties of emergence. Paper presented at the Social Agents: Ecology, Exchange, and Evolution Conference, Chicago. http://www.soc.surrey.ac.uk/staff/ngilbert/ngpub/paper148_NG.pdf (2002)

Hall, N., Paul, D.: Causation and Counterfactuals. MIT Press, Cambridge (2004)

Hanson, J., Crutchfield, J.: Computational mechanics of cellular automata: an example. Phys. D **103**, 169–189 (1997)

Holland, J.: Hidden Order. How Adaptation Builds Complexity. Addison-Wesley, New York (1995)

Holland, J.: Emergence. From Chaos to Order. Basic Books, New York (1998)

Hovda, P.: Quantifying weak emergence. Mind. Mach. **18**, 461–473 (2008)

Humphreys, P.: How properties emerge. Philos. Sci. **64**, 53–70 (1997)
Humphreys, P.: Synchronic and diachronic emergence. Mind. Mach. **18**, 431–442 (2008)
Huneman, P.: Combinatorial vs. computational views of emergence: emergence made ontological? Philos. Sci. **75**, 595–607 (2008a)
Huneman, P.: Emergence and adaptation. Mind. Mach. **18**, 493–520 (2008b)
Huneman, P.: Computer sciences meet evolutionary biology: issues in gradualism. In: Torres, J.L., Pombo, O., Symons, J., Rahman, S. (eds.) Special Sciences and the Unity of Science, pp. 200–225. Springer, Dordrecht (2011)
Huneman, P.: Determinism and predictability: lessons from computational emergence. Synthese **185**(2), 195–214 (2012)
Huneman, P.: Mapping an expanding territory: computer simulations in evolutionary biology. Hist. Philos. Life Sci. **36**(1), 60–89 (2014)
Israeli, N., Goldenfeld, N.: On computational irreducibility and the predictability of complex physical systems. Phys. Rev. Lett. **92**, 074105 (2004)
Jonas, J.: The Phenomenon of Life: Toward a Philosophical Biology. Harper & Row, New York (1966)
Kant, I.: Critique of Judgment. Hackett, Indianapolis (1987 [1790])
Kim, J.: Making sense of emergence. Philos. Stud. **95**, 3–36 (1999)
Klee, R.: Microdeterminisms and concepts of emergence. Philos. Sci. **51**, 44–63 (1984)
Kuorikoski, J., Lehtinen, A., Marchionni, C.: Robustness analysis disclaimer: please read the manual before use! Biol. Philos. **27**, 891–902 (2012)
Langton, C.: Artificial life. In: Langton, C. (ed.) Artificial Life. SFI studies in the sciences of complexity, Proc. Vol. VI. Addison-Wesley, Redwood City (1989)
Laughlin, R.: A Different Universe: Reinventing Physics from the Bottom Down. Basic Books, New York (2005)
Laughlin, R.B., Pines, D., Schmalian, J., Stojkovi, B., Wolynes, P.: The middle way. Proc. Natl. Acad. Sci. U. S. A. **97**(1), 32–37 (2000)
Levins, R.: The strategy of model building in population biology. In: Sober, E. (ed.) Conceptual Issues in Evolutionary Biology, 1st edn, pp. 18–27. MIT Press, Cambridge, MA (1966)
Lewis, D.: Causation. J. Philos. **70**, 556–567 (1973)
Lipton, P.: Inference to the Best Explanation. Routledge, London (1991)
Matthewson, J., Weisberg, M.: The structure of trade-offs in model building. Synthese **170**(1), 169–190 (2009)
Mc Laughlin, B.: The rise and fall of British emergentism. In: Beckermann, A., Flohr, H., Kim, J. (eds.) Emergence or Reduction? de Gruyter, Berlin (1992)
Nagel, K., Rasmussen, K.: Traffic at the edge of chaos. In: Brooks, R. (ed.) Artificial Life IV. MIT Press, Cambridge, MA (1994)
Nagel, T.: What is it like to be a *Bat*? Philos. Rev. **83**(4), 435–450 (1974)
O'Connor, T.: Emergent properties. Am. Philos. Q. **31**, 91–104 (1994)
Orzack, S.H., Sober, E.: A critical assessment of Levins's the strategy of model building in population biology (1966). Q. Rev. Biol. **68**, 533–546 (1993)
Piaget, J.: La construction du réel chez l'enfant. Delachaux et Niestlé, Paris (1937)
Rasmussen, S., Baas, N., Mayer, B., Nilsson, M., Olesen, M.: Ansatz for dynamical hierarchies. Artif. Life **7**(4), 329–353 (2002)
Roe, S.A.: Matter, Life, Generation. Eighteenth-Century Embryology and the Haller-Wolff Debate. Cambridge University Press, Cambridge (1981)
Salzberg, C., Antony, A., Sayama, H.: Genetic diversification and adaptation of self-replicators discovered in simple cellular automata. In: Proceedings of the Sixth International Conference on Humans and Computers (HC-2003), pp. 194–199. University of Aizu, Aizuwakamatsu (2003)
Sayama, H.: Spontaneous evolution of self reproducing loops in cellular automata. In: Bar-Yam, Y., Minai, A.A. (eds.) Unifying Themes in Complex Systems Volume II: Proceedings of the Second International Conference on Complex Systems, pp. 363–374. Westview Press (1998)
Schelling, T.: Models of segregation. Am. Econ. Rev. **59**(2), 488–493 (1969)

Seager, W.: Emergence and efficacy. In: Erneling, C., Johnson, D. (eds.) The Mind as a Scientific Object Between Brain and Culture, pp. 176–192. Oxford University Press, Oxford (2005)
Sellars, W.: Philosophy and the scientific image of man. In: Colodny, R. (ed.) Frontiers of Science and Philosophy, pp. 35–78. University of Pittsburgh Press, Pittsburgh (1962)
Shalizi, C., Haslinger, R., Rouquier, J.B., Klinkner, C., Moore, C.: Automatic filters for the detection of coherent structures in spatiotemporal systems. *ArXiv CG/0508001* (2006)
Silberstein, M.: Reduction, emergence and explanation. In: Silberstein, M., Machamer, P. (eds.) Blackwell Guide to the Philosophy of Science, pp. 80–107. Blackwell, Oxford (2002)
Tassier, T.: A model of fads, fashions and group formations. Complexity **9**(5), 51–61 (2004)
Weisberg, M.: Robustness analysis. Philos. Sci. **73**, 730–742 (2006)
Wilson, J.: Non-reductive physicalism and degrees of freedom. Br. J. Philos. Sci **61**(2), 279–311 (2010a)
Wilson, R.: The third way of agent-based social simulation and a computational account of emergence. J. Artif. Soc. Soc. Simul **13**(3), 8 (2010b). http://jasss.soc.surrey.ac.uk/13/3/8.html
Wimsatt, W.: False models as means to truer theories. In: Nitecki, N., Hoffman, A. (eds.) Neutral Models in Biology, pp. 23–55. Oxford University Press, Oxford (1987)
Wimsatt, W.: Aggregation: reductive heuristics for finding emergence. Philos. Sci. **64**, S372–S384 (1997)
Wolfe, C., Normandin, S.: Vitalism and the Scientific Image in Post-Enlightenment Life Science, 1800–2010. Springer, Dordrecht (2013)
Woodward, J.: Making Things Happen. Oxford University Press, New York (2003)

Chapter 23
A Comparison of the Semantics of Natural Kind Terms and Artifactual Terms

Luis Fernández Moreno

Abstract This paper aims to compare the semantics of natural kind terms and that of artifactual terms. To that end, we rely on the natural kind terms' theory regarded as paradigmatic in contemporaneous semantics, the one put forward by Putnam, who sketched the extension of the semantics of natural kind terms to artifactual terms. In this paper we develop such extension concerning the reference of artifactual terms, although the reference fixing theory we advocate differs from that of Putnam's. On the other hand, we propose a view on the meaning of these terms which conflicts with the one it would follow from extending to such terms Putnam's view of meaning on natural kind terms.

Keywords Natural kind terms • Artifactual terms • Semantics • Reference fixing • Meaning

23.1 Introduction

Putnam's main target in (1975b), a classic of contemporaneous semantics, is to criticize the "traditional theory of meaning" on *natural kind terms* and to propose an alternative view of the semantics of such terms. Nevertheless, in the section "Other words" of (1975b), he sketches the extension of the semantics of natural kind terms to other types of terms, mainly other sorts of general terms, including especially *artifactual terms*[1] (and socio-legal terms which I will leave aside).

It is appropriate to begin by pointing out that, although the function of natural kind terms is to refer to natural kinds, and the one of artifactual terms to artifactual kinds, the distinction between natural kinds and non-natural kinds – especially artifactual kinds – runs into some difficulties because it does not coincide with

[1]I will understand the expression "artifactual term" as interchangeable with "artifactual kind term" and consequently "artifact" with "artifactual kind".

L. Fernández Moreno (✉)
Complutense University of Madrid, Madrid, Spain
e-mail: luis.fernandez@filos.ucm.es

J. Redmond et al. (eds.), *Epistemology, Knowledge and the Impact of Interaction*,
Logic, Epistemology, and the Unity of Science 38, DOI 10.1007/978-3-319-26506-3_23

the distinction between kinds whose members are found in nature and those whose members are man-made. Among the former are the kinds mud, dust and shrub, which are not considered as natural kinds, while amongst the latter are the kinds technetium and diamond, which are plausibly regarded as natural kinds (see LaPorte 2004, p. 18). On this matter I agree with the view that the distinction between natural kinds and non-natural kinds "is not *sharp*, but rather one of *degree*, so that perhaps kinds can ultimately be classified into *more* or *less* natural ones along a spectrum of some sort, with clear cases on either side and a good bit of indeterminacy in the middle" (Koslicki 2008, p. 203; see LaPorte 2004, p. 23).

There are still two groups of terms that in contemporary semantics have been regarded as prototypical natural kind terms.[2] They are terms for biological kinds, like "cat" and "tiger" – including terms for botanical kinds, as "elm" and "beech" – and terms for natural materials and in particular for chemical substances, such as "water" and "gold". In this paper, I will mainly take into consideration the latter sort of natural kind terms, since they are the ones most frequently mentioned by Putnam in (1975b). Such type of terms has been denominated by many authors, Putnam among others, *substance terms* (or names),[3] and I will usually employ this terminology in the following. Nevertheless, since Putnam's semantics concerning substance terms constitutes a particular case, however central, of his theory regarding natural kind terms, I will frequently take up a stance at this more general level, since what I will assert regarding substance terms would fundamentally apply to the rest of natural kind terms.

23.2 Putnam's Extension of the Semantics of Natural Kind Terms to Artifactual Terms

Putnam characterizes the traditional theory of meaning in different ways in his papers (1970) and (1975b). If we take into consideration some features of that theory in these papers (see Putnam 1970, pp. 139 f. and 1975b, p. 242), and restricting for now our considerations to natural kind terms, we could distinguish two versions of it. According to one of them, namely, the *conjunction*-of-properties version, the meaning of a natural kind term is given by a conjunction of properties – and the term is defined through them –, so that the conjunction of *all* those properties determines the reference or extension of the term. In this regard, it is noteworthy that Putnam identifies, as I will do in the following, the reference of a general term, like natural kind terms and artifactual terms, with its *extension*, i.e., with the set of entities to which the term applies. According to the other version of the traditional theory of meaning concerning natural kind terms, the meaning of a natural kind term is given

[2] In (1973) Putnam distinguishes between physical magnitude terms and natural kind terms, although he claims that there are semantic similarities between both sorts of terms. In this paper I will not deal with physical magnitude terms.

[3] See, for instance, Putnam (1975b, p. 231) and Putnam (1990, p. 58).

by a *cluster* of properties – and the term is defined by that cluster – so that a *sufficient number* of properties in the cluster determine the reference of the term.

Putnam assumes that the properties to which the traditional theory of meaning of natural kind terms resorts are exclusively the properties that according to our common sense beliefs characterize the paradigmatic members of the natural kind – the normal distinguishing properties. In Putnam's terminology, we can differentiate two sorts of these properties, the *semantic markers* and the *stereotype*. The difference between them is that the former are the most central or hardly revisable of such properties; however, in order to simplify my considerations, up to Sect. 23.4, I will leave aside that distinction, talking simply of the *stereotype* of the term. In any case, Putnam sustains that those properties are not analytically associated with the natural kind term in question and do not determine its reference. Putnam emphasizes two contributions involved in the reference determination of natural kind terms: the contribution of the environment and that of the society.[4] On the one hand, the extension of a natural kind term depends on how *our* environment or world is, since this is determined by *underlying properties* of the members of the natural kind belonging to our world. On the other hand, the elucidation of these properties is the object of scientific research, and those who carry it out or use its results, i.e., the *experts*, will have a better knowledge than average speakers of the membership conditions into the extension of a natural kind term. In this regard there is a *division of linguistic labor*, according to which the reference of natural kind terms, as they are used by average speakers, depends on the reference of such terms in their use by experts, since the former are willing to defer to the latter their judgments concerning the membership of an entity into the corresponding natural kind.

Come to this point, Putnam proposes two ways of fixing the reference of a natural kind term like "water" in a speaker's idiolect. One of them is an *ostensive definition* while the other lies in a *description* which Putnam denominates *operational definition*.[5] Regarding the first procedure, let us suppose that a speaker, in the presence of another one, points to a glass containing water and utters the following ostensive definition "This liquid is water". The force of such "definition" is that a sample of substance is a sample of water if and only if it is a sample of the *same* actual liquid as *this* is a sample of. On the other hand, by means of an "operational definition" of the term "water" Putnam understands a (definite) description formed with the general terms that express the properties included in the stereotype of that term: the liquid that is colorless, transparent, tasteless, thirst-quenching, etc. (in the actual world). The force of such "definition" is that a sample of substance is a sample of water if and only if it is a sample of the *same* actual liquid whose paradigmatic samples in *our* world, i.e. in the *actual* world, possess such (normal distinguishing) properties.

[4]On these contributions see Putnam (1975b, pp. 227–234, 245, 265 and 271), as well as (1988), chapter 2.

[5]Concerning the first procedure, see (1975b, pp. 225 and 229 ff.), and on the second, (1975b, pp. 229 f. and 232 f.).

An important aspect shared by both definitions is that the fixing of the reference of natural kind terms involves – explicitly or implicitly – the use of *indexicals*, like "this", "our" or "actual". Putnam alludes to this feature asserting that natural kind terms possess "an *indexical component*" (Putnam 1975b, p. 234 and 1988, p. 33) – although he sometimes expresses himself in a shorter, but rather misleading way when he affirms that natural kind terms are indexical (Putnam 1975b, ibid.) –, and claims that this feature implies that those terms are *rigid designators*. In this regard Putnam holds that a natural kind term like "water" is rigid because it applies, with respect to all possible worlds, only to samples that share with the samples of water in the *actual world* the same underlying properties, i.e., the same nature or essential properties. Thus, a sample of substance in any possible world is a sample of "water" if and only if it is a sample of H_2O.

Let us now turn to the section "Other words" of (1975b), which begins as follows:

> The points made [concerning natural kind terms] apply to many other kinds of words as well [. . .] Let us consider [. . .] the names of artifacts – words like 'pencil', 'chair', 'bottle', etc. (Putnam 1975b, p. 242).[6]

As already said, in Putnam's characterization of the traditional theory of meaning for natural kind terms, but also applicable to artifactual terms, two versions can be distinguished, the conjunction-of-properties version and the cluster version. Putnam characterizes that theory concerning artifactual terms in the following way:

> The traditional view is that these words [such as 'pencil', 'chair', 'bottle', etc.] are certainly defined by conjunctions, or possibly clusters, of properties. Anything with all of the properties in the conjunction (or sufficiently many of the properties in the cluster, on the cluster model) is necessarily a *pencil*, *chair*, *bottle*, or whatever. In addition, some of the properties in the cluster (on the cluster model) are usually held to be *necessary* (on the conjunction-of-properties model, *all* of the properties in the conjunction are necessary). *Being an artifact* is supposedly necessary, and belonging to a kind with a certain standard purpose – e.g. 'pencils are artifacts' and 'pencils are standardly intended to be written with' are supposed to be necessary. Finally, this sort of necessity is held to be *epistemic* necessity – in fact, analyticity. (Putnam 1975b, p. 242).

In this regard, Putnam proposes the following epistemic thought experiment. He asks us to imagine that the pencils of our world were discovered to be not artifacts but organisms. If it is epistemologically possible that pencils are organisms, the property of being an artifact is not analytically associated with the term "pencil". However, let us assume that our beliefs about pencils on Earth are right, i.e., that they are artifacts standardly intended to be written with, while on Twin Earth (conceived as a possible world different from the actual world) its inhabitants have the same beliefs concerning the things that they called "pencils", but later it is discovered that these "pencils" are organisms. In this case, Putnam claims that we should assert that there are no pencils on Twin Earth. The justification of this claim is that artifactual terms, like natural kind terms, are *rigid designators*; hence, if pencils in

[6] I will speak of artifactual terms or artifactual kind terms – see note 1 above – instead of names of artifacts.

our environment – i.e., on the Earth – are artifacts, pencils are also artifacts in all possible worlds in which they exist. Thus, the statement that pencils are artifacts will be metaphysically necessary, although not epistemically necessary. Putnam asserts:

> When we use the word 'pencil', we intend to refer to whatever has the same *nature* as the normal examples of the local pencils in the actual world. 'Pencil' is just as *indexical* as 'water' or 'gold'. (Putnam 1975b, p. 243).

Putnam is using the term "pencil" as a prototypical example of an artifactual term; thus, he intends that his claims about that term apply to all artifactual terms, and especially to *everyday* artifactual terms – i.e., terms for everyday (objects which are) artifacts – like the examples he mentions ("chair" and "bottle", besides "pencil"). This sort of artifactual terms is the one I will focus on in the following.[7]

In order to examine Putnam's extension of the semantics of natural kind terms to artifactual terms, we will have to deal with the semantics of both sorts of terms, more precisely, with their reference and their meaning. Let us begin by their reference.

23.3 The Reference of Natural Kind Terms and Artifactual Terms

In a reference theory it can be distinguished between a theory of reference fixing and a theory of reference borrowing (or transmission). In this regard I will initially accept that the former sort of reference theory regarding natural kind terms and artifactual terms is similar, at least to the extent that both terms can be introduced by ostension to paradigmatic members of the kinds or by description.[8] As already said, according to Putnam there are two ways of fixing the reference of a natural kind term – we will leave aside Putnam's qualification that this holds for a speaker's idiolect. One of them is an *ostensive definition* while the other is a *description* or, in Putnam's words, an *operational definition*. Let us now concentrate on the first one.

Since the ostension by itself is *ambiguous* it has to be supplemented, at least, by a general term – at large by a descriptive component. Thus, let us bear in mind the corresponding clause of the ostensive definition of the term "water" proposed by Putnam, i.e., "This liquid". An expression of this type is denominated a complex demonstrative, and also a demonstrative description (see Abbott 2010, p. 6). Nevertheless, to our end it will be suitable to consider this kind of expressions as a sort of *indexical descriptions* – this is Putnam's terminology in (1988); see below.

[7] There is not a precise delimitation criterion for everyday artifactual terms, but the examples mentioned by Putnam are clear cases of them.

[8] As it is well known, Kripke proposes in the second lecture of (1980) a theory of reference fixing for proper names that he extends in the third lecture to natural kind terms, according to which a term is introduced in an *initial baptism* in which its reference is fixed by ostension or "by a description" (Kripke 1980, pp. 96 f.). These procedures of reference fixing are similar to those proposed by Putnam concerning natural kind terms.

These are descriptions which contain an indexical, be it or not a demonstrative, and regardless whether they begin or not with that indexical.

Nevertheless, since the ostensive contact with members of the kind involves a *causal* component, an appropriate theory of reference *fixing* by ostension for natural kind terms and artifactual terms is *descriptive-causal*. The question to be posed is how much *descriptive* content is to be included in the descriptions involved in that sort of reference fixing.

In this respect we should attend to a problem pointed out by Devitt and Sterelny (see 1999, pp. 90 ff.). These authors have alleged that the introduction by ostension of natural kind terms involves us in the *qua*-problem. This problem is double or, as I will also say, it has two parts. The first problem arises because an object belongs to different *sorts* of kinds; let us imagine, for example, the introduction by ostension of the term "gold" by pointing to a gold ring; this object belongs to the natural kind gold but also to the artifactual kind ring. For this reason, the introducer of a natural kind term will have to associate with the term some description that classifies the term to be introduced as a natural kind term. The source of the second part of the *qua*-problem is that the entities pointed to for the introduction of a natural kind term will belong to different *natural kinds*; thus, for instance, a sample of gold belongs to the kind gold, but also to the kind metal, to the kind element, etc. Since the entities involved in the introduction of a natural kind term will share many underlying properties or "natures", it is required to pick out the one relevant to the reference of the natural kind term. Thus, Devitt and Sterelny have proposed that in order to solve the *qua*-problem involved in the reference fixing of natural kind terms the introducer of the term should associate, consciously or unconsciously, with the term two sorts of descriptions:

> First some description that in effect classifies the term as a natural kind term; second, some descriptions that determine which nature of the sample is relevant to the reference of the term. (Devitt and Sterelny 1999, p. 92).

In this regard it is noteworthy that in a writing posterior to (1975b) Putnam proposes the following "indexical description" – his words – for the reference fixing of substance terms: "stuff that behaves like and has the same composition as *this*" (Putnam 1988, p. 38). Concerning this description I should make two comments. First, the term in question will designate a (sort of) stuff and, more precisely, a (sort of) substance[9]; therefore, the term will be a substance term and hence a natural kind term; thus, the first part of the *qua*-problem is sorted out. Second, in that description

[9]Putnam adds that the indexical description in question is uttered by "someone who is 'focusing' on a particular sample of substance" (Putnam 1988, p. 38). On this matter, Putnam asserts that he has taken the notion of "focusing" from Alan Berger; see Putnam (1988, pp. 33 and 130, n. 14). It is remarkable that although in that footnote 14 Putnam claims that Berger introduces the notion of "focusing" in *Terms and Truth* (Cambridge, Mass.: MIT Press, 1988), the published version of that book appeared later, that is, Berger (2002); thus, it is to be assumed that Putnam had access to a preliminary manuscript of that book. Concerning Berger's notion of "focusing" see chapters 1 and 2 of Berger (2002).

it is alluded to the behaviour as well as to the composition of a sample of substance, but since its behaviour will be explained by its composition, the *description* of that behaviour would contribute to solving the second part of the *qua*-problem: the composition of the sample responsible for such and such behaviour is the "nature of the sample [. . .] relevant to the reference of the term", in Devitt's and Sterelny's words, and this composition will not be shared by all metals, by all elements, and so on. Thus, Putnam's indexical description includes the components to sort out the two parts of the *qua*-problem. However, we can propose a slightly different indexical description to fix the reference of natural kind terms which will also solve the (two parts of the) *qua*-problem concerning the sort of natural kind terms substance terms are: the indexical description "This substance with such and such behaviour".

In order to make a proposal concerning the reference fixing of artifactual terms, we have to take as a starting point a definition of artifact. A prototypical definition of the notion of artifact, which I will assume henceforth, is the following: an artifact is "an object that has been intentionally made for some purpose" (Hilpinen 2011, p. 1). However, this definition is also fulfilled by some entities regarded as belonging to natural kinds, since some members of natural kinds are products of farming and breeding; thus, the distinction between artifacts and some natural entities is not sharp – as we conceded in Sect. 23.1; see also Soavi (2009, pp. 10 f.). Nevertheless, our considerations will focus on examples of objects that are *clearly* artifacts, and as already said, everyday artifacts, i.e., artifacts of everyday use.

The people who make artifacts, among whom there are speakers who fixed the reference of the corresponding terms, are usually denominated the "makers" or "designers", but I will opt for the latter expression. In this regard, it is plausible to claim that the original designers that fixed the reference of an artifactual term by ostension were also involved in the *qua*-problem. Thus, on the one hand, those designers had to associate, consciously or unconsciously, with the term a description that classifies the object as an artifact – let us remember the example of the gold ring mentioned above. The description could be simply "This artifact". On the other hand, such designers had to specify the artifactual kind in question, since an artifact can be a member of several artifactual kinds, e.g., a chair is a member of the artifactual kind chair, but also of the artifactual kind furniture, etc.

For this reason, we will have to elucidate the properties involved in the identity and individuation of an artifact, since they will determine the reference or extension of the corresponding term. In this regard, it is appropriate to take into account Schwartz's view, who considers artifactual kinds as a sort of nominal kinds, and he makes a contraposition between natural kinds and nominal kinds, which are respectively the extensions of natural kind terms and nominal kind terms:

> Nominal kind terms differ from natural kind terms in that the extension of a nominal kind term is not gathered by an underlying trait. Perhaps the best examples of nominal kind terms are the names of common artifacts such as 'pencil', 'bottle', and 'chair' [these are just the examples of artifactual terms mentioned by Putnam in (1975b)]. The extension of a

> nominal kind term *is* determined by an analytical specification of superficial features such as phenomenal properties, and/or form, function, or origin. (Schwartz 1980, p. 182).[10]

Leaving aside, until we reach Sect. 23.4, the claim about the analytical character of the features involved in the determination of the extension of artifactual kind terms, and restricting our considerations to the properties mentioned by Schwartz in this passage, I would like to make some remarks. The *function* (or purpose) for which an artifact has been intentionally made – let us say, for short, its *intended function* – will play, according to the definition of an artifact we have accepted, an outstanding role in the determination of the extension of an artifactual term, but it cannot be the only property involved in that determination, since, for instance, a chair, an armchair and a sofa can share the same intended function, i.e., to be things to sit on. Before attending to additional properties, it is noteworthy to point out that the intended function with which an original designer made an artifact may not coincide with the intended function for which it is regularly used at present. Although in case of conflict between those functions I would opt for the present one, we should indicate that different sorts of designers can be recognized. As Grandy asserts:

> If a kind of artifact has a function that was not intended by the *original designer*, then someone else recognized that possibility and so we should simply broaden our criterion and recognize that in many cases a kind of object has *multiple designers*/creative users. (Grandy 2007, p. 28; emphasis added).

Concerning other properties that can contribute to the determination of the extension of artifactual terms, besides the intended function – or functions –, which will always be involved, I regard as acceptable the other properties mentioned by Schwartz. I will allude to the "phenomenal properties and/or form" of an artifact as its *appearance* (see Abbott 1989), and I will also include into them the (perceptible) physical properties of the members of the artifactual kind in question. By the *origin* of an artifact I will understand the way in which it has been made, and this feature will involve – unlike Putnam's claim – that it is not plausible, or even possible, to imagine discovering that in our world pencils are organisms.

Two other properties that can be taken into account are the way in which an artifact performs its intended function, to which I will allude as its *manner of use*, as well as (features of) its *internal structure* – see Losonsky (1990). In this regard I claim that the resort to the latter may be relevant in some cases for the determination of the reference of the corresponding artifactual term, since the internal structure of an artifact will generally contribute to the performance of its intended function; thus, for instance, the fact that pencils have a lead standardly made of graphite, which as a rule allows to distinguish pencils from pens. Accordingly, among the properties that determine the reference of artifactual terms my proposal is to include

[10]Schwartz's view in (1978) was somewhat different, since he claimed that "we can give an analytical specification in terms of form and function of what it is to be a member of the nominal kind" (Schwartz 1978, p. 572).

as the fundamental property of the corresponding artifacts their intended function, but there can also be taken into account their appearance, origin, manner of use and internal structure, these last sorts of properties having a different weight or relevance for the reference determination of various everyday artifactual terms. In this regard my proposal, unlike Putnam's, is a version of the "traditional theory" of the reference fixing of artifactual terms, the cluster version: the reference of an artifactual term is fixed by a sufficient number of the cluster of the mentioned properties, although the most important property is the intended function. However, come to this point, I would like to introduce two caveats. First, the five mentioned properties, as I have already done concerning the internal structure of pencils, are to be qualified with the adjective *standard* or with the adverb *standardly*, as Putnam did in the passage quoted above (Putnam 1975b, p. 242). Second, since those properties will characterize the (paradigmatic) members of an artifactual kind in *our* world, i.e., in the *actual world*, the general terms that express the mentioned properties will be understood as containing implicitly the clause "in the actual world".

In order to develop the extension of the semantics of natural kind terms to artifactual terms, we should propose a type of "indexical description" to fix the reference of everyday artifactual terms which will also sort out the (two parts of the) *qua*-problem. The indexical description I propose would adopt the following form: a/an T – here the artifactual term – is "*this* artifact with such and such intended function, such appearance, such origin, such manner of use and such (features of its) internal structure". This indexical description can be the basis for a descriptive-causal theory of reference fixing for artifactual terms.[11] However, if we consider that the five mentioned properties of an artifact are the properties included in the stereotype of the corresponding artifactual term – in the broad sense of stereotype we are assuming up to Sect. 23.4. –,[12] we can also put forward a sort of "operational definition" concerning artifactual terms, substituting in the mentioned indexical description the demonstrative "this" by the definite article "the", and this operational definition implicitly contains indexicals, to wit the term "actual", since according to the proposal made at the end of the last paragraph, the general terms which express such properties are to be understood as implicitly containing the clause "in the actual word".

If we accept that the reference of artifactual terms is fixed by either type of description, or rather by a sufficient number of the properties expressed through the general terms that appear in those descriptions, which explicitly or implicitly

[11] I say "descriptive-causal" instead of merely "descriptive" because of the causal component involved in the ostensive contact with members of the kind.

[12] By the stereotype of an artifactual term there should be understood the properties that according to our common sense beliefs characterize the paradigmatic members of the artifactual kind. Concerning everyday artifacts every competent speaker in the use of the corresponding terms knows the intended function, the appearance and the manner of use of such artifacts. Although it should be recognized that the knowledge of their origin and features of their internal structure can be more imperfect, these are, as a rule, easily discernible.

contain indexicals, we will agree with Putnam's claim that those terms have an indexical component and they are therefore *rigid designators*.

In a passage quoted above from Putnam (1975b, p. 243) he alluded to the *nature* of the paradigmatic pencils in the actual world. In this regard, the "nature" or essential properties of the members of natural kinds and artifactual kinds are different. The essential properties of the members of a natural kind are their underlying properties, while according to my proposal the essential properties of the members of an artifactual kind are, besides the property of being an artifact, which is taken for granted, their intended function, although the other four properties mentioned above could also be regarded as essential in case they should necessarily be linked to that intended function.

Come to this point we can take into consideration the theory of reference *borrowing* for natural kind terms and artifactual terms. In this regard the main question to be posed is whether for speakers to be reference borrowers of a kind term they have to associate with the term some properties or descriptions that determine its reference. In the case of natural kind terms, it is usually claimed that this is *not* the case, because the properties that most speakers associate with a natural kind term are those included in the stereotype of the term, but these properties do not determine the reference of a natural kind term, and the underlying properties which do are usually, at most, only known by experts in Putnam's sense.

Concerning artifactual terms we have to distinguish at least two cases. Although they do not refer to everyday artifacts, there are artifactual terms subject to the division of linguistic labour (see Kornblith 1980, p. 113 and 2007, pp. 43 f.), and regarding the latter the descriptions that most speakers associate with them, if any, would be very imprecise; for instance, in the case of the artifactual term "cyclotron" the description could be an indefinite one, like "a sort of machine" or a similar one, which puts a very slight restriction on the reference of the term. Concerning everyday artifactual terms – the artifactual terms on which we are focusing our considerations –, like "pencil", most speakers learn them, and borrow their reference, in ostensive contact with samples of the kind, but in this case I would claim that to be able to be involved in the reference borrowing most speakers have to associate with those terms a *cluster* of properties including, besides the intended function, *some* of the other mentioned properties (mainly appearance and manner of use of such artifacts, but in some cases also their origin and features of their internal structure), and that a sufficient number of the properties in this cluster determines their reference. Thus, with respect to everyday artifactual terms and to most speakers I advocate for a *descriptive-causal* theory of reference borrowing.

I can summarize my foregoing considerations concerning the reference of artifactual terms in the following way. An indexical description (or in Putnam's terminology in 1975b, an ostensive definition) as well as a definite description constituting an "operational definition" can fix the reference of everyday artifactual terms; thus a descriptive-causal theory as well as a purely descriptive theory can be adequate for the reference fixing of that sort of artifactual terms, but in either case the reference of the term is fixed by a sufficient number of the properties (in the cluster) expressed by the general terms that appear in those descriptions. However, since

most speakers learn those terms, and borrow their reference, in ostensive contact with samples of the kind, I advocate for a descriptive-causal reference theory for reference borrowing. Lastly, the reference fixing for artifactual terms subject to the division of linguistic labour could be purely descriptive or descriptive-causal, but their reference borrowing is fundamentally historical-causal.[13]

23.4 The Meaning of Natural Kind Terms and Artifactual Terms

The semantics of a term does not, however, only include its reference but also its *meaning*. Since there are different notions of meaning I should indicate which will be the one I will take into consideration; in this regard I will assume an epistemic notion of meaning, according to which by the meaning of a term I will understand the components required to be known by a speaker for him to be a competent user of the term.

Come to this point, we could ask about the meaning of natural kind terms and of artifactual terms. In Putnam's semantics, which we have relied on for our considerations, the meaning of a natural kind term is given by means of a finite sequence. In (1975b, p. 269) Putnam asserts that the members of the sequence that would constitute the meaning of the term "water" would include the following ones: syntactic markers (mass noun, concrete), semantic markers (natural kind, liquid), stereotype (colorless, transparent, tasteless, thirst-quenching, etc.), and extension: H_2O (give or take impurities).

With respect to these components I would like to make the following remarks. First, in writings posterior to (1975b) Putnam has changed his view concerning the last component:

> Once we have discovered the chemical composition of water in the actual world to be H_2O [. . .] we do not call any other actual or hypothetical substance 'water' unless it is *similar in composition* to this. But 'similar in composition' is a somewhat vague notion. (1983, p. 63; Putnam's emphasis).

Thus, the description of the extension of the term "water", which describes one of the components of the sequence that delivers the meaning of that term, should be H_2O or a composition similar to it (give or take impurities). Second, for two speakers to have learnt the meaning of the term "water" sameness of stereotype is not required "but rather sufficient similarity, where what counts as 'sufficient' is highly context sensitive" (Putnam 1987, p. 271). Third, in order for a speaker to be competent in the use of a natural kind term Putnam requires that the speaker should *only* know the stereotype, the semantic markers and the syntactic markers of the

[13]The theory for reference borrowing put forward by Kripke in (1980) for proper names and natural kind terms is historical-causal.

term, although the knowledge by most competent speakers of at least part of those components of the meaning of the term will be implicit. Thus, if it is explained to the speaker what is a mass term (or noun), a concrete term and a natural kind term, the speaker would assent to the fact that the term "water" has the first two features and that the substance water has the third property.

According to the view of the meaning of a term that I have assumed, I agree with the proposal that the meaning of a natural kind term is given by these three components, one of them being the properties of the stereotype (or properties sufficiently similar to them), and therefore I do *not* accept Putnam's thesis according to which the meaning of a natural kind term by itself determines its extension. In Putnam's case that thesis is trivially true, since the extension of a term is one of the components of its meaning.

Concerning the meaning of everyday artifactual terms I would propose a *similar* view, again leaving aside the extension, but I would not talk of a finite sequence of elements, but of a *cluster* of them – in fact, I am inclined to sustain the same thesis with regard to the meaning of natural kind terms –, which includes three properties: the syntactic markers, the semantic markers and the stereotype. As the syntactic markers of an artifactual term I propose the features of being a general term and a concrete term.[14] The semantic markers should be the property of being an artifact and the intended function of the artifact (or a function sufficiently similar to it). The stereotype includes some of the following properties: appearance, manner of use, origin, and features of the internal structure (or properties sufficiently similar to them). But in the case of artifactual terms, a sufficient number of the properties in the cluster that constitutes their meaning *do* determine their extension.

There is still a question to be posed, i.e., whether any of those properties are analytically associated with an everyday artifactual term. In this regard, I would claim that it could be held that the intended function is analytically associated with the term, but if this thesis could be put into question by rather unimaginable epistemic thought experiments, a more plausible thesis, and in any case the one I support, is that the inclusive disjunction of the properties in the cluster that constitutes the meaning of an artifactual kind term is analytically associated with the term – obviously from the first thesis the second would follow. Putnam alleged that the property of being an artifact is not analytically associated with the term "pencil", since – he claimed – we could imagine it was discovered that pencils are organisms – although that claim could be objected to by resorting to the origin of pencils, i.e., the way they are made. However, Putnam did not put into question the analytical association with the term "pencil" of the property that pencils are standardly intended to write with, nor did he take into consideration other properties like their appearance, manner of use, origin, or features of their internal structure.

It is indeed difficult to propose epistemic thought experiments through which we would imagine that we are mistaken concerning many of the properties included in the cluster that constitutes the meaning of an artifactual term, although concerning

[14] Whereas not all artifacts are concrete entities, everyday artifacts are.

pencils Nelson proposed a thought experiment of this sort, where "the ubiquitous objects called 'pencils' are actually devices scattered about Earth by malevolent aliens who by such means manipulate human activity" (Nelson 1982, p. 362), and where pencils do not have most of the properties that we associate with them. However, if my position is acceptable it can be replied that if we were to make an object that should satisfy the cluster of properties whose inclusive disjunction I have considered as analytically associated with the corresponding artifactual term, the object in question would be an artifact of that kind, even if we could imagine discovering that we were wrong concerning the properties we have regarded as possessed by paradigmatic members of the artifactual kind (see Schwartz 1983, p. 477 f. for a similar sort of response). Another more radical reply, which Schwartz would not endorse, is that the people who manufacture pencils – and many educated people – know that the situation imagined by Nelson – like the one imagined by Putnam – is *impossible*: philosophers sometimes pretend to "imagine" all sorts of things, including impossible things.[15]

We can conclude by stating an important difference between the semantics of natural kind terms and artifactual terms: an entity belongs to the extension of a natural kind term if it possesses the underlying properties of the paradigmatic members of the kind, which are not analytically associated with the term, while an entity belongs to the extension of an artifactual kind term if it possesses a sufficient number of the properties in the cluster that constitutes the meaning of the term, and the inclusive disjunction of those properties is analytically associated with the term. Therefore, the semantic theory I advocate with respect to artifactual terms is one of the versions of the "traditional theory": the cluster theory.[16]

References

Abbott, B.: Nondescriptionality and natural kind terms. Linguist. Philos. **12**, 269–291 (1989)
Abbott, B.: Reference. Oxford University Press, Oxford (2010)
Baghramian, M. (ed.): Reading Putnam. Routledge, London/New York (2013)
Berger, A.: Terms and Truth. Reference Direct and Anaphoric. MIT Press, Cambridge, MA (2002)
Burge, T.: Some remarks on 'externalisms'. In: Baghramian (2013), pp. 263–271 (2013)
Devitt, M., Sterelny, K.: Language and Reality. An Introduction to the Philosophy of Language. MIT Press, Cambridge, MA. 2nd ed., rev. and extended, 1st ed., 1987 (1999)
Grandy, R.E.: Artifacts: parts and principles. In: Margolis, E., Laurence, S. (2007), pp. 18–32 (2007)
Hilpinen, R.: Artifact. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy (Winter 2011 Edition) http://plato.stanford.edu/archives/win2011/entries/artifact/ (2011)
Kornblith, H.: Referring to artifacts. Philos. Rev. **89**, 109–114 (1980)

[15]This is also a charge brought against Putnam's thought experiment of Twin Earth in Putnam (1975b); see Burge (2013, p. 269 f.) and Putnam (2013, p. 274).

[16]This paper has been supported by the Spanish Ministry of Economy and Competitiveness in the framework of the research project FFI2014-52244-P.

Kornblith, H.: How to refer to artifacts. In: Margolis, E., Laurence, S. (2007), pp. 138–149 (2007)
Koslicki, K.: The Structure of Objects. Oxford University Press, Oxford (2008)
Kripke, S.: Naming and Necessity. Blackwell, Oxford (1980)
LaPorte, J.: Natural Kinds and Conceptual Change. Cambridge University Press, Cambridge (2004)
Losonsky, M.: The nature of artifacts. Philosophy **65**, 81–88 (1990)
Margolis, E., Laurence, S. (eds.): Creations of the Mind. Theories of Artifacts and Their Representation. Oxford University Press, Oxford (2007)
Nelson, J.A.: Schwartz on reference. South. J. Philos. **20**, 359–365 (1982)
Putnam, H.: Is semantics possible? In: Kiefer, H., Munitz, M. (eds.) Languages, Belief and Metaphysics. State University of New York Press, New York. Reprinted in Putnam (1975a), pp. 139–152 (1970)
Putnam, H.: Explanation and reference. In: Pearce, G., Maynard, P. (eds.) Conceptual Change. Reidel, Dordrecht. Reprinted in Putnam (1975a), pp. 196–214 (1973)
Putnam, H.: Mind, Language and Reality. Philosophical papers, vol. 2. Cambridge University Press, Cambridge (1975a)
Putnam, H.: The meaning of 'meaning'. In: Putnam (1975a), pp. 215–271 (1975b)
Putnam, H.: Realism and Reason. Philosophical papers, vol. 3. Cambridge University Press, Cambridge (1983)
Putnam, H.: Meaning holism and epistemic holism. In: Cramer, K., et al. (eds.) Theorie der Subjektivität, pp. 251–277. Suhrkamp, Frankfurt (1987)
Putnam, H.: Representation and Reality. MIT Press, Cambridge (1988)
Putnam, H.: Realism with a Human Face. (Edited and with introduction by J. Conant). Harvard University Press, Cambridge, MA (1990)
Putnam, H.: Comments on Tyler Burge. In: Baghramian (2013), pp. 272–274 (2013)
Schwartz, S.P.: Putnam on artifacts. Philos. Rev. **87**, 566–574 (1978)
Schwartz, S.P.: Natural kinds and nominal kinds. Mind **89**, 182–195 (1980)
Schwartz, S.P.: Reply to Kornblith and Nelson. South. J. Philos. **21**, 475–479 (1983)
Soavi, M.: Realism, Antirealism and Artefact Kinds. CLEP, Padova (2009)

Chapter 24
Models, Representation and Incompatibility. A Contribution to the Epistemological Debate on the Philosophy of Physics

Andrés Rivadulla

Abstract Theoretical models play a fundamental role in the methodology of theoretical physics. There is no branch in contemporary physics, whether it be cosmology, astrophysics or microphysics, where such models are not used.

Theoretical models are idealized constructs of a single phenomenon or about a limited empirical domain. They are intended to both save the phenomena and to make testable predictions about the domain they are concerned with. In any case, models are not susceptible to being true or verisimilar representations of certain aspects of reality. On this point, I disagree with most realist philosophers of science.

Models make use of extant theories and are of particular use in domains lacking theories. Moreover, in a historical sequence of theoretical models about a certain domain, not every model is compatible with previous ones. This is the case of Ptolemaic and Copernican cosmological models or of Einsteinian and Newtonian gravitational models. The incompatibility among models (and even theories) about the same domain is the most serious issue facing standard convergent realism. In order to illustrate this situation, I am going to focus on various kinds of theoretical models employed by nuclear physics.

Keywords Theoretical models • Newtonian mechanics • Nuclear physics • Theoretical incompatibility • Scientific realism

Complutense research Group 930174 and research project FFI2014-52224-P financed by the Ministry of Economy and Competitiveness of the Kingdom of Spain. I am very grateful to an anonymous referee for comments on an earlier version of this paper.

A. Rivadulla (✉)
Departamento de Lógica y Filosofía de la Ciencia, Universidad Complutense de Madrid, Madrid, Spain
e-mail: arivadulla@filos.ucm.es

J. Redmond et al. (eds.), *Epistemology, Knowledge and the Impact of Interaction*, Logic, Epistemology, and the Unity of Science 38, DOI 10.1007/978-3-319-26506-3_24

24.1 Introduction

Since the development of mathematical physics, theoretical models have played an increasingly significant role in the methodology of science, and have become indispensable to a better understanding of how theoretical physics deals scientifically with Nature. Even in ancient astronomy, the geometrical models of the Universe provided the means to *save* the *erratic* movements of celestial bodies.

The fruitfulness of the use of models is nowadays evident. There is no branch in contemporary theoretical physics where models are not employed, whether it is in astrophysics, cosmology or microphysics. Moreover, statistical mechanics provided models of gases for quantum systems such as bosons and fermions. Kepler's astronomy supplied an idealized geometrical model of planetary motion, and Newtonian mechanics provided a highly successful model of gravitational phenomena, one which continued to be accepted until it was substituted by the more successful model provided by relativity theory – the pseudo-Euclidean four-dimensional Minkowski spacetime. Hydrodynamics offered a model of the Universe as a fluid of galaxies in relativistic cosmology, etc.

Theoretical models are idealized constructs of a single phenomenon or about a limited empirical domain. Models make use of extant theories and *are of particular use in domains where there are no theories available*. Following theoretical models must be coherent with the already accepted theoretical background. The common feature of all theoretical models is that they are intended to save observed phenomena and to provide empirically testable predictions in their domains, and in any case neither verisimilitude, nor isomorphism, nor similarity, reflect all possible ontological relationships between models and the world. Thus they are not susceptible to being true or verisimilar representations of aspects of reality. The basic reason for this is that the fundamental requirement of any model is empirical success, and since we do not know the phenomenon under investigation *as it really is*, the inference from success to verisimilitude cannot be legitimate. Weaker demands like similarity are even less justified. Thus we are not allowed to claim that models *represent* the phenomena themselves. Neither as representation nor as simulations of reality do models relate to Nature. This is the point at which I disagree with realist philosophers of science in general.

A further problem with the realist picture of scientific progress is that in a historical sequence of theoretical models of a certain domain, not every model is a compatible extension of previous ones, as it is the case with Copernican, Keplerian and Newtonian celestial models, or with the continental drift and plate tectonic model in geophysics. Indeed, models of the sequence can sometimes be mutually incompatible, like Ptolemaic and Copernican celestial models, or Einsteinian and Newtonian gravitational models – free floating in geometric curved spaces models of gravity, against models of forces acting at a distance. Indeed, they may contradict or deny each other, as is the case with, for example, contracting Earth and expanding Earth models in geophysics, or independent particles and collective nuclear models in microphysics. These examples of incompatibility, and even of contradiction,

between models in the earth sciences, gravitational physics and microphysics, all indicate the Achilles' heel in scientific realism, and show that theoretical models are intrinsically fallible constructs intended to deal predictably with Nature, that they are not more or less faithful representations of an independent reality. The fact that a model presumably refers to something out there in the world does not constitute any cogent reason for claiming that it represents the outside world, since the only access we have to the world is mediated throughout the model itself. Thus the model cannot at the same time be both judge and part of the 'cognitive' task.

There is thus no sense in the claim that models represent aspects of the world in a realist sense of the term *representation.* The reasonableness of a model choice is provided only by the *predictive balance*, i.e. the weighting of the empirically tested predictive power of the competing models.

In accordance with the concept that theoretical models are non-representational, I claim that neither does it make any sense to affirm that models can explain, unless under the *theoretical explanation* of a physical construct we merely conceive of the fact that the *explanandum* – be it a fact, a phenomenological law, a theoretical law, or even a theoretical model or a theory – can be mathematically deduced within the framework of another physical construct of higher theoretical level. For this kind of non-metaphysical theoretical explanation, only coherence with the theoretical background is needed, but without any representational requirements.

24.2 Theoretical Models and the Epistemological Debate

It is relatively easy to agree that theoretical physics is empirically very successful. The recent discovery in July 2012 at the CERN in Geneva of a new particle compatible with the Higgs boson confirms that theoretical physics is highly capable of achieving empirical success. The question is if it is also full of truth as well.

To begin with I am going to present several conflicting viewpoints among major contemporary physicists on the role and possibilities of theoretical physics. For instance, Steven Weinberg (1998: 48) claims that "What drives us onward in the work of science is precisely the sense that there are truths out there to be discovered, truths that once discovered will form a permanent part of human knowledge." But contradicting this view Stephen Hawking affirms (EL PAIS: 13.04.2005) that:

> Una teoría es tan sólo un modelo matemático para describir las observaciones, y no tiene derecho a identificarse con la realidad, sea lo que sea lo que esto signifique. Podría ser que dos modelos muy diferentes lograran describir las mismas observaciones: ambas teorías serían igualmente válidas, y no se podría decir que una de ellas fuera más real que la otra.[1]

[1] A theory is only a mathematical model for describing observations, and it does not have the right to be identified with reality, whatever that means. It may happen that two very different models are successful in describing the same observations: both these theories will be equally valid, and it would not be possible to state that one of them was any more real than the other.

Along the same lines as Weinberg, Lee Smolin (2007: 7) claims that "Physics should be more than a set of formulas that predict what we will observe in an experiment; it should give a picture of what reality *is*." He states furthermore that "realism provides the motivation driving most scientists." (Smolin, op. cit.: 9). On the contrary, Paul Dirac (1963) considers that "If the physicist knows how to calculate results and compare them with experiment, he is quite happy if the results agree with his experiments, and that is all he needs." And describing Niels Bohr's perspective on the current state of the quantum mechanical picture of the world, Penrose (1989:226) claims that "Quantum theory ... provides merely a calculation procedure, and does not attempt to describe the world as it actually 'is'."

This controversy is a clear example of the philosophical realism/antirealism debate, translated into the realm of theoretical physics. In order to deal with this, I shall focus primarily on the role that theoretical models play in mathematical physics. Why focus on models and not on theories? There are many reasons:

1. Not all theories are models, but all models are theories (Popper 1994)
2. Theories can be very complex entities. They are not the best candidates (insofar as they are theoretical constructs) for tackling questions of representation and truth.
3. Theories are not always available in every research domain. But theoretical models are.
4. Theoretical models are particularly important in domains that are lacking theories. For instance in nuclear physics:

> No complete theory exists which fully describes the structure and behaviour of complex nuclei based solely on knowledge of the force acting between nucleons. However, great progress has been and is being made with the aid of conceptual models designed to give insight into the underlying physics of the inherently complex situation. (Lilley 2007: 35)

Since I shall focus on theoretical models, it seems appropriate to provide an answer to the question of what kind of entities the theoretical models of physics are. According to Popper (1994: 172) "it seems to be quite unavoidable in the construction of models, both in the natural and in the social sciences, that they oversimplify the facts, and thus do not represent the facts truly". And J. S. Lilley (2007: 35) asserts:

> A model embodies certain aspects of our knowledge and, almost invariably, incorporates simplifying assumptions which enable calculations to be made. A successful model should be able to give a reasonable account of the properties it was designed to address and also make predictions of other properties which can be checked by experiment.

Moreover, Eisberg and Resnick (1974: 591) claim that "A model provides a description of only a limited set of phenomena, without regard to the existence of contrary models used for the description of other sets".

But, what kind of entities are the theoretical models of physics? In my view, the theoretical models of physics satisfy the following criteria, which I shall postulate from the outset, i.e. without any direct justification and from an anti-realist viewpoint, but which I hope will be justified in the course of the following

pages. It is nonetheless necessary to state here that, since I assume the physicist's viewpoint of the role and function of theoretical models in contemporary physics, I do not endorse the so-called semantic view defended by a great many philosophers of science. Thus my theses on theoretical models are that:

1. Theoretical models are hypothetical constructs intended to both save the phenomena and to make testable predictions about the empirical domains with which they are concerned.
2. Models are idealized constructs of a phenomenon or about a limited empirical domain.
3. They assume the form of a mathematical equation or of a series of closely related equations.
4. They are particularly useful in disciplines lacking a theory, for instance in stellar astrophysics and nuclear physics.
5. Theoretical models must inescapably be consistent with both already accepted laws of physics and with available empirical data.
6. The condition sine qua non for the acceptance of a theoretical model is *empirical success.*

Since I am tackling the issue of realism in contemporary philosophy of physics, it also seems advisable to present *standard scientific realism* as the epistemological doctrine according to which both contemporary mature theories are (at least) approximately true and the theoretical terms they employ refer empirically. The strongest argument on behalf of scientific realism is Putnam-Boyd's *no-miracle argument*. Presented by Hilary Putnam (1975:73) as "The positive argument for realism is that it is the only philosophy that doesn't make the success of science a miracle", and in (1978: 18): "the typical realist argument against idealism is that it makes the success of science a *miracle*", this argument states, according to Richard Boyd (1984: 43), that "If scientific theories weren't (approximately) true, it would be miraculous that they yield such accurate observational predictions."

Most contemporary scientific realists rely on realism on the basis of an *optimistic meta-abduction*, an *inference to the best explanation*. As a matter of fact, in a typically abductive manner Paul Thagard (1988:150) argues on behalf of scientific realism in the following way [Compare with Peirce (1965: *CP* 5.189)] :

1. Truth is seen as a property of scientific theories.
   [The surprising fact, *C,* is observed]
2. But to accept realism is to suppose that scientific theories can be said to be true.
   [But if *A* were true, *C* would be a matter of course]
3. There is no reason not to see it [i.e., *truth*, A, R.] as a property of metaphysical theories such as realism.
   [Hence, there is reason to suspect that *A* is true]

Now, since it is practically impossible for a good argument to exist in the philosophy of science without an excellent counter-argument, scientific realism typically encounters Laudan's *pessimistic meta-induction* (1981). In several steps:

1. "*there can be* (and have been) *highly successful theories some central terms of which are non-referring*" (p. 226) "*a part of the historical success of science has been success exhibited by theories whose central terms did not refer.*" (p. 226)
2. "*a realist would never want to say that a theory was approximately true if its central terms failed to refer.*" (p. 230)
3. "a theory may be empirically successful even if it is not approximately true" (p. 244)

Stathis Psillos (1999: 101) has given a standard formulation of Laudan's pessimistic meta-induction: "The history of science is full of theories which at different times and for long periods had been empirically successful, and yet were shown to be false ... Therefore, ... our current successful theories are likely to be false."

Since it is reasonable to assume that (a) neither any terms of our contemporary theories do refer empirically, and (b) nor do our own contemporary theories need to be true in order to be empirically successful, then *truth seems not to be necessary in science*. This view raises the question of whether or not the role played by theoretical models in physics does indeed support scientific realism.

24.3 The Case of the Newtonian Celestial Model and the Incompatibility Argument

Newtonian mechanics offered the most enduring theoretical model of the Universe until the advent of relativity theory. It is based on the Law of Universal Gravitation, and results in the following achievements, among many others:

1. Kepler's 3rd Law, as interpreted in the framework of Newtonian mechanics, allows for the calculation of the mass of any star whatsoever, for instance the Sun's mass, which amounts to 1.989×10^{30} kg.

   It also permits the calculation of the mass included in the orbit of the Sun, which is located in Arm Orion of the Galaxy, around the centre of the Galaxy, if we know that the radius of the Sun's orbit is 8.5 kpc, and the orbital period is 2.4×10^{8} years. This mass value is about 9.4×10^{10} solar masses. (cf. Martínez et al. 2005: 208)
2. The expression of the intensity of the gravitational field of Earth allows for the calculation of the mass of Earth: 5.9×10^{24} kg.
3. The hypothesis of the existence of black holes: stars whose escape velocity – due to their mass – is faster than the speed of light. In 1783 John Michell (1724–1793) referred to them for the first time as *dark stars.*
4. Even the critical density of the Universe can be calculated in the framework of Newtonian mechanics etc.

Do these achievements support the view that the Newtonian celestial model is verisimilar, or that there is some probability of its being true? If they did support such a view, then Newtonian mechanics should be able successfully to tackle new

theoretical challenges: for example, the advancement of the planets' perihelia, the deflection of light by the Sun, and the existence of black holes.

It is true that the Newtonian celestial model was by no means silent when confronted with these challenges. Indeed, a possible answer to the case of Mercury's 'anomalous' perihelion might be provided in *Newtonian approximation*, although this value falls far below the observed value. As to the deflection of photons by the Sun, the Newtonian model gives $0''.87$ of arc, a result that was obtained by Georg von Söldner (1776–1883). Finally, and also in Newtonian approximation, the redshift z of a photon trying to escape the gravitational field of a star amounts to $z = 1$.

Contradicting these predictions, General Relativity Theory (GRT) provides the following results (cf. Rivadulla 2004a, notes 18 and 19):

1. In the case of Mercury, the calculated value is $43''.03$/century, which conforms well to the observed perihelion's advance value.
2. The value of the light deflection by the Sun predicted by GRT amounts to $1''$,75 of arc, as first observed by Arthur Eddington in 1919.
3. Since the gravitational redshift of a photon is, according to GRT, $z = \infty$, this conforms very well to the intuition of a *black hole*.

In conclusion, the Newtonian gravitational model fails where GRT is successful. Indeed, Popper (1994:172) might well ask: "Can any model be true? I do not think so. Any model, whether in physics or in the social sciences, must be an over-simplification, it must omit much, and it must overemphasize much." Moreover,

> models are always and necessarily somewhat rough and schematic over-simplifications. Their roughness entails a comparatively low degree of testability. For it will be difficult to decide whether a discrepancy is due to the unavoidable roughness or to a mistake in the model. (Popper 1994:170)

Along the same lines, in relation to the Newtonian celestial model, Popper (1994:172) points out:

> Take a Newtonian model of the solar system. Even if we assume that Newton's laws of motion are true, the model would not be true. Though it contains a number of planets – in the form, incidentally, of mass-points, which they are not – it contains neither the meteorites nor the cosmic dust. It contains neither the pressure of the light of the sun nor that of cosmic radiation. It does not even contain the magnetic properties of the planets, or the electric fields which result in their neighbourhood from the movements of these magnets. And – perhaps the most important – it does not contain anything representing that action of the distant masses upon the bodies of the solar system. It is, like all models, a vast over-simplification.

Although Popper cannot be blamed for ignoring the problems of dark matter and dark energy, he is allegedly suggesting that, if the Newtonian solar model did take into account everything it leaves out, then it might be a faithful representation of reality. But this is incorrect. We do not even know that what the Newtonian model still preserves – a more or less efficient application of the Gravitational Law among the masses of the Solar system – actually reflects an underlying reality which is structured in the way established by Newtonian mechanics.

What guarantee do we have that what keeps objects falling near the Earth's surface, and the planets orbiting around the Sun, and the satellites around their own planets, are the effect of an attractive force that is directly proportional to the product of their masses (taken in twos) and inversely proportional to the square of the distance between them (again taken in twos)? We have no such guarantee, since we do not even know that these gravitational forces really exist, or whether gravitation itself is the result of different circumstances. Indeed, it might be, as GRT claims, that gravitation is merely the effect of the geometry of spacetime. Thus the explanation of gravitational phenomena would have no relationship with physical forces.

To conclude: Popper's claim that models greatly oversimplify cannot guarantee that what models preserve even minimally mirrors reality. The entities postulated in those super-simplified models could easily not exist. In the same way that epicycles, equants, deferents, eccentrics, and the whole paraphernalia of both Ptolemaic entities and Aristotelian spheres do not exist, Newtonian gravitational forces and potentials might also be inexistent. If this were the case, then Newtonian mechanics would not reflect reality even minimally, whether over-simplified or not.

There is also no guarantee that Relativity Theory itself is true or truth-like. As Arthur Fine (1984: 92) points out:

> I believe the majority opinion among working, knowledgeable scientists is that general relativity provides a magnificent organizing tool for treating certain gravitational problems in astrophysics and cosmology. ... For relativistic physics, then, it appears ... that most who actually use it think of the theory as a powerful instrument, rather than as expressing a 'big truth'.

This suggests that, from a non-realist point of view, Truth plays no role in science. In my view, theory is not the house of truth, and theoretical thinking does not harbour the Truth, although among many theoretical physicists optimism regarding the possibilities of physics has not vanished. For instance, Steven Weinberg (2001: 206) maintains:

> in recent years we have seen electrodynamics and the theories of other forces in nature merge into the modern Standard Model of elementary particles. We hope that in the next great step forward in physics we shall see the theory of gravitation and all of the different branches of elementary particle physics flow together into a single unified theory. This is what we are working for and what we spend the taxpayers' money for. And when we have discovered this theory, it will be part of a true description of reality.

[Nonetheless this optimism at the same time contrasts with the frustration caused by the failure of physics to extend our knowledge of the basic laws since 1975 – roughly the period referred to by Weinberg:

> we have failed. ... For more than two centuries, until the present period, our understanding of the laws of nature expanded rapidly. But today, despite our best efforts, what we know for certain about these laws is no more than what we knew back in the 1970s. (Smolin 2007: Introduction)]

What, then, is the fate of standard scientific realism? Realist philosophers of science assume that scientific progress is somehow linear. Even when they accept the existence of scientific revolutions as rational processes due to the presence of

continuity elements – meaning existence of limiting cases –, they do not seriously take into account the existence of inter-theoretical incompatibilities. Incompatibility is omnipresent in the realm of theoretical physics:

- incompatibility at the level of theoretical entities that mutually deny one another.
- incompatibility at the level of fundamental postulates: between Copernicus and Ptolemy, but also between Einstein and Newton, whereby the picture of the four-dimensional pseudo-Euclidean Universe contradicts that of the Newtonian three-dimensional Euclidean Universe.
- incompatibility between Relativity Theory and Quantum Mechanics.
- incompatibility between different current Quantum Theories: determinist and indeterminist, linear and non-linear.
- incompatibility between background-dependent and background-independent theories (string theories and quantum-gravity theories), etc.

Incompatibility contradicts the realist idea of convergence to truth. Thus from the viewpoints of Copernicus and Einstein, the theories of Ptolemy and Newton cannot respectively be true or close to the truth. It becomes clear that what theory, theoretical thinking, harbours is not truth, but incompatibility.

To support my viewpoint, let me cite Eugene Wigner (1967/1995: 234–235/546–547), Nobel laureate, 1963:

> We now have, in physics, two theories of great power and interest: the theory of quantum phenomena and the theory of relativity. These two theories have their roots in mutually exclusive groups of phenomena. Relativity theory applies to macroscopic bodies, such as stars. (...) Quantum theory has its roots in the microscopic world ... The two theories operate with different mathematical concepts – the four dimensional Riemann space and the infinite dimensional Hilbert space, respectively. So far, the two theories could not be united, that is, no mathematical formulation exists to which both of these theories are approximations. All physicists believe that a union of the two theories is inherently possible and that we shall find it. Nevertheless, it is possible also to imagine that no union of the two theories can be found.

And in 1995: 591 Wigner notes:

> As far as the consistency of present day physics is concerned, I have serious reservations. Though quantum mechanics has been successfully applied to the determination of many macroscopic constants, its ultimate validity for macroscopic systems is not clear. [...] The full body of quantum mechanics, as applied to macroscopic systems can not be completely verified – a conclusion one arrives at very reluctantly ... The general theory of relativity appears to represent the opposite extreme.

On his side Lee Smolin (op. cit.: 4–5) claims:

> There is no way we can have two theories of nature covering different phenomena, as if one had nothing to do with the other. ... In the atomic realm, where quantum physics reigns, we can usually ignore gravity. (...) The other realm is that of gravitation and cosmology. In that world, we can often ignore quantum phenomena.

As applied to the epistemological debate, I interpret this situation as the recognition of the failure of standard scientific realism.

24.4 The Case of Nuclear Models

24.4.1 Microscopic, Single-Particle or Independent-Particle Nuclear Models vs. Macroscopic or Collective Models

Macroscopic or collective models conceive of the atomic nucleus as a fluid, where the only interesting movement is the collective movement of the nucleons. The most representative collective nuclear model is the *liquid-drop model*. According to this, the atomic nucleus is supposed to look like an incompressible liquid drop.

The most impressive prediction of this nuclear model is Weizsäcker's Semi-Empirical Mass Formula (SEMF), which allows for the computation of the mass of nuclei, accounts for their stability, predicts the binding energy of nuclei, predicts the atomic number Z of the most stable nuclei with mass number A ($A = Z + N$), etc.

Nonetheless the nuclear liquid drop model does not work for nuclei with small mass number A. Moreover there are nuclei with certain values of Z (number of protons) and/or N (number of neutrons) that are unusually stable and are not predicted by SEMF. These are nuclei with nucleon numbers 2, 8, 20, 28, 50, 82 and 126, called *magic numbers*. Their behaviour is not predicted by the SEMF based on the liquid-drop model. In terms of van Fraassen's (2008) representation theory, the theoretical model SEMF does not represent correctly the data model provided by the experimental values of the binding energy per nucleon (cf. Lilley 2007: Figure 2.3).

In order to tackle the problems presented by the collective liquid-drop model, physicists decided to abandon it and to assume that any nucleon in a nucleus interacts with, or experiences, an average field due to the other nucleons.

The shell model is the most successful nuclear model that conceives of nuclei as being formed of closed shells of nucleons. (cf. Rivadulla 2004b:148–151) Each shell or sub-shell has a nuclear orbit associated with it. This model allows us to account for the existence of nuclear *magic numbers*. All nuclei with closed shells, and nuclei with *magic numbers,* are symmetrically spherical, i.e. the cuadrupolar electric moment $Q = 0$ – Q being a measure of the degree of the deviation from sphericity – and they are very stable. *Double magic nuclei* ^{4}He, ^{16}O, ^{40}Ca, ^{90}Zr and ^{208}Pb are of course spherical as well, and they are most stable. Moreover the single-particle model predicts with great accuracy the behaviour of nuclei with odd A.

Again, in spite of being very successful in terms of prediction, shell models have several shortcomings: (1) Shell models fail in the prediction of the magnetic dipolar moment of nuclei with odd A. Their failure is due to the incorrect assumption that the nuclear magnetic dipolar moment was that of the unpaired nucleon. But not all nucleons are always paired, so that their total angular and magnetic dipolar moments cancel each other out. (2) Moreover the shell model also fails to accurately predict the electric cuadrupolar nuclear moments Q_i.

24.4.2 *Collective Models, Vibrational and Rotational models, vs. Independent-Particle Models*

Contrary to nuclei situated in the neighbourhood of magic numbers, which are practically spherical, i. e. $Q \approx 0$, nuclei with $Q < 0$ or $Q > 0$ are not. To account for the discrepancies of the predictions of Q in the single-particle model with observations, nuclear physicists return to collective models which follow the image of the nuclear liquid-drop model. Two new collective models enter the stage: vibrational and rotational models.

The *vibrational model* describes the vibrations around the spherical form of light nuclei ($A < 150$). According to the values of the cuadrupolar (and even octupolar) electric moment, the atomic nucleus oscillates between prolate ($Q > 0$) and oblate ($Q < 0$) forms, passing through the spherical form. (cf. Lilley 2007; Ferrer Soria 2006).

Far from the regions of magic numbers, nuclei are deformed even in their fundamental state, i. e. they deviate from sphere about 20 %. Typically we find them among the *lantanides* ($150 \leq A \leq 190$) and the *actinides* ($220 \leq A \leq 250$); they are known as *deformed nuclei* and they are described by so called *rotational models* (cf. Lilley 2007; Ferrer Soria 2006).

24.5 Conclusion

Given the evident incompatibility existing among the different models in nuclear physics, and in general in different parts of theoretical physics, is it reasonable to believe that models truly or faithfully represent those parts of Nature they are concerned with? Endorsing Richard Rorty's (1980: 377–378) anti-Platonism, I affirm that the theoretical realm is not the home of Truth, that theoretical thinking does not harbour the Truth, and that Theory does not mirror Nature. And if it does, whenever that might be, we cannot know it.

References

Boyd, R.: The current status of scientific realism. In: Leplin, J. (ed.) Scientific Realism. University of California Press, Berkeley (1984)

Dirac, P.: The evolution of the physicist's picture of nature. Sci. Am. **208**(5), 45–53 (1963)

Eisberg, R., Resnick, R.: Quantum Physics of Atoms, Molecules, Solids, Nuclei, and Particles. Wiley, New York (1974)

Ferrer Soria, A.: Física nuclear y de partículas. Universitat de Valencia, Valencia (2006)

Fine, A.: The natural ontological attitude. In: Leplin, J. (ed.) Scientific Realism. University of California Press, Berkeley. Reprinted in A. Fine, The Shaky Game. Einstein, Realism and the Quantum Theory. University Press, Chicago, 1986. Reprinted in D. Papineau, The Philosophy of Science. University Press, Oxford, 1996 (1984)

Laudan, L.: A confutation of convergent realism. Philos. Sci. **48**, 19–49 (1981)
Lilley, J.S.: Nuclear Physics. Principles and Applications. Wiley, Chichester (2007)
Martínez, V.J., et al.: Astronomía Fundamental. Universitat de Valencia, Valencia (2005)
Peirce, C.: Collected Papers. Harvard University Press, Cambridge, MA (1965)
Penrose, R.: The Emperor's New Mind. University Press, Oxford (1989)
Popper, K.: The Myth of the Framework. In Defence of Science and Rationality. Routledge, London (1994)
Psillos, S: Scientific Realism. How Science Tracks Truth. Routledge, London (1999)
Putnam, H.: Philosophical Papers, vol. I. University Press, Cambridge (1975)
Putnam, H.: What is realism. In: Putnam, H. (ed) Meaning and the Moral Sciences. Routledge, London (1978)
Rivadulla, A.: The Newtonian limit of relativity theory and the rationality of theory change. Synthese **141**, 419–429 (2004a)
Rivadulla, A.: Éxito, Razón y Cambio en Física. Un enfoque instrumental en teoría de la ciencia. Trotta, Madrid (2004b)
Rorty, R.: Philosophy and the Mirror of Nature. University Press, Princeton (1980)
Smolin, L.: The Trouble with Physics. The Rise of String Theory, the Fall of a Science, and What Comes Next. Houghton Mifflin Company, Boston-New York (2007)
Thagard, P.: Computational Philosophy of Science. The MIT Press, Cambridge, MA (1988)
Van Fraassen, B.: Scientific Representation. Paradoxes of Perspective. Clarendon, Oxford (2008)
Weinberg, S: The revolution that didn't happen. N. Y. Rev. Books. **XLV**(15), 48–52 (1998)
Weinberg, S.: Facing Up: Science and Its Cultural Adversaries. Harvard University Press, Cambridge, MA (2001)
Wigner, E.: The unreasonable effectiveness of mathematics in the natural sciences. In: Symmetries and Reflections, pp. 222–237. Indiana University Press, Bloomington. Reprinted in E. P. Wigner. Philosophical Reflections and Syntheses, pp. 534–549. Springer, Berlin 1995 (1967)
Wigner, E.: Physics and its relation to human knowledge 1977. In: Wigner, E.P. (ed) Philosophical Reflections and Syntheses, pp. 584–593. Springer, Berlin (1995)

Chapter 25
Fictions in Legal Science: The Strange Case of the Basic Norm

Juliele Maria Sievers

Abstract If the presence of fiction in natural sciences is sufficiently known and accepted, the same doesn't seem to be the case when it comes to legal science. The presence of fictions in Law is unquestioned and can be traced since Roman law, even if its legitimacy remains a matter of great divergence among critics. However, the legitimacy of the use of fictions by natural sciences or Philosophy is attested by famous examples of thought experiments, for instance. Considering this context, we will analyze the use of fictions made by a special kind of science dealing with the regulation of our behavior, namely legal science.

Our aim is to analyze the use of fiction by the legal science under the light of the legal theory proposed by Hans Kelsen (1881–1973), especially concerning his proposal that the legitimization of the whole positive legal system is based on a fiction, called the Basic Norm (*Grundnorm*). The difference, we sustain, is that this "norm" must be seen as a methodological or scientific tool, and not as an ordinary norm among others in the legal system. We will try to elucidate how can a fiction display such an important function and still preserve the "principle of purity" of the kelsenian legal theory.

Keywords Fictions • Hans Kelsen • Basic Norm • Legal Science

25.1 Legal Science: Meaning and Particularities

It is a central aspect of the kelsenian legal philosophy the fact that the Law, namely the overall set of norms, must be clearly separated from the science that describes and studies this set. Another differentiation cherished by Kelsen separates neatly this kind of normative science, which takes the norm as its object, from the natural

This text is the outcome of the work as a Post-Doc researcher in the Universidade Federal de Santa Maria and funded by the program PNPD-CAPES.

J.M. Sievers (✉)
Philosophy Department, Universidade Federal de Santa Maria, Santa Maria, Brazil
e-mail: julisievers@yahoo.com.br

J. Redmond et al. (eds.), *Epistemology, Knowledge and the Impact of Interaction*,
Logic, Epistemology, and the Unity of Science 38, DOI 10.1007/978-3-319-26506-3_25

sciences, which takes the facts of nature as its object. These two ideas constitute the methodological frame in which Kelsen formulates his Pure Theory of Law. Let's understand at what length the notion of "purity" of Law deals with each one of these two kelsenian paradigms.

A common mistake made by some of Kelsen's contemporaries was not to distinguish Law from legal science.[1] Certainly, one should concede that the content of Legal Science, in the context of legal Positivism, can sometimes be very confusing. The theory of Law describes the norms of a particular normative system.[2] But the point is that the norm described by the science has no longer its normative power or, in the terms of Kelsen, it is no longer a valid norm. It is exactly the fact of the validity of the norm that is being described, and a fact cannot (legally) compel anyone.

But let us start from the beginning, by explaining the definition of the term "legal norm" according to Kelsen's positivist theory.

The definition of "norm" is linked to the concept of a command or an order, with the important detail that this order must come from an authorized person, as an expression – mostly in the form of an imperative – of a will. Since this *will* must come from a person authorized by the Law itself, there is always a strict relation between the legal production and the legal power.[3] The norm is the meaning of an objective act of will, coming from an authorized person. It is marked by the presence of the "ought" particle,[4] meaning that we're not in the domain of the "Is", but in the domain of the "Ought" (*Sollen*). About this duality, Kelsen says[5]:

> *When someone commands or prescribes, he wills that something ought to happen. The Ought – the norm – is the meaning of a willing or act of will, and – if the norm is a prescription or command – it is the meaning of an act directed to the behavior of another person, an act whose meaning is that another person (or persons) is to behave in a certain way*.

According to Bobbio,[6] to enact a norm is always to be able to do so, and the authorization comes from no one but the Law itself. This shows the importance of the fact that the procedures to menage the legal norms and handle the legal affairs must always come from the Law itself, according to the principle of "Purity".[7]

---

[1]This "confusion" is linked to the denial of the duality of "Is" and "Ought". For a review on several classical cases, see KELSEN, H. *General theory of Norms*, Oxford University Press 2011 [1979], pp. 63–82.

[2]Cf. Spaak, T. "Kelsen and Hart on the Normativity of Law". In: *Perspectives on Jurisprudence: Essays in Honour of Jes Bjarup*. Peter Wahlgren, ed., pp. 397–414, 2005.

[3]Cf. van Roermund, B. "Authority and Authorization". In: *Law and Philosophy*, Springer, Vol. 19, No. 2, pp. 201–222, Mar. 2000.

[4]Even if the norm in question is not mandatory, but concerning permissions, empowerment or derogation, the "ought" element is always preserved.

[5]Kelsen, H. General Theory of Norms. Oxford University Press, New York 2011 [1979], p. 2.

[6]Bobbio, N. *Teoría General del Derecho*. Santa Fé de Bogotá – Colombia: Editorial Temis, 1997.

[7]What is implicit in this conception is that there is no place to logical treatment in the inner domain of Law, i. e., in the production or interpretation (decision) of legal norms.

25.1.1 Some Examples

So what is needed to create a positive legal norm? First of all, there must be a prior, more general norm to support the existence of the new one. Then, the subject must be an authorized person capable of enacting an objective act of will. Those elements are necessary to characterize the norm as a valid one: its specific existence in the frame of a legal order; the fact of its connection to a legal system *via* a more general norm, and the compulsoriness that links up all its addressees. This late obligatory aspect is in fact the objectiveness of the act of will coming from the Judge/Legislator. A *subjective* act of will consists only in a command, without legal force, it means the expression of a personal will towards a specific case, and that's not what Law is about. Instead, the *objective* act of will comes only from an authorized, impartial, neutral person, and the meaning of this objective act of will is the legal norm.[8]

Kelsen illustrates the latter by giving the example of a gangster's demanding for money.[9] When the gangster asks you to give him all your money, this order has not the same meaning as when the tax officer asks you for the money, namely, that the person towards whom the order is formulated *ought* to render a determined amount of money. The tax officer's order is actually a binding valid norm, because it is based on the Law, and the officer in question has the authority which was given by the Law to perform in the way he does. The gangster's order represents a subjective act of will, and it has the meaning of a command, but not of a valid norm.[10]

While the command is the expression of a desire, the norm is the expression of a duty, of an "ought" (*Sollen*). The notion of "ought" introduces us to the separation between the legal prescriptive field of the norms – the domain of the "ought" – and the factual descriptive field of the legal science – the domain of the "is".

Concerning the relations between the science and its scientific object – here, legal science and Law itself – the English philosopher of Law Herbert Hart[11] uses a nice example to illustrate the fact that, even if the legal science deals with valid norms, it doesn't have any normative legal power.[12] He uses the example of the

[8]By this approach, an objective command is not only the psychic event of the expression of a will. This can be seen in the case of a testament, for instance. In a valid testament, the subjective act of will of the person in question obtains its objectivity trough the Law: once it is legally legitimated, the command of the person in question will remain beyond his own existence, when he will no longer be able to express his will. This demonstrates the independency of the compulsoriness of the command from the subjective act of will.

[9]This example can be found in various passages of KELSEN, H. "Théorie Pure du Droit", 2e traduction par Ch. Eisenmann, Dalloz, Paris 1962.

[10]Actually, what turns this norm into a binding norm is the fact of assuming the existence of the Basic Norm in relation to such a normative system. We come to this issue later in this text.

[11]HART, H. "Visita a Kelsen". In: *Lua Nova*. No. 64. São Paulo Jan./Apr. 2005 [1963].

[12]This aspect is essential to our future analysis of the Basic Norm as an element of legal science, and not of Law itself.

relation between someone who speaks a foreign language and his, let's say, English interpreter. If a German captain in a concentration camp says out loud to his English or American prisoners "*Stehen Sie auf*!", the interpreter will probably also say out loud the words "Stand up!". The interpreter will do his best to show, by his intonation maybe, or by the expression in his face, that what the Captain said was not a begging or a simple request: it was an order. The point is: how do we must consider the sentence "Stand up!" in respect of its original in German? Is it a second order? Is it the same order? Is it the emission of an order? Well, the interpreter has no authority to emit orders. His job is to interpreter the orders of the captain and, if the order is obeyed or not, it was the Captain who was obeyed or disobeyed. Hart then says that, in perfect accordance to Kelsen's theory, what happens in such case is that the captain's order is being described by the interpreter, that the "Stand up!" was an order in a descriptive sense, that the original imperative was used in a descriptive, not prescriptive, sense.

Lastly, we must note that the great distinction between the legal norm and the statement of the legal theory lies in the fact that the statement can be said to be true or false, while the norm can only be said to be valid. Actually, even the facts themselves cannot be valued as true or false, but only as existing or not existing: true or false are only the statement made about those facts. The predicates true/false can be said only in relation to statement of the "is" domain, to the descriptions of the facts which can be made by any science. But they are not linked to the very object of these sciences, as the natural facts, for instance. The norm is exactly the object being described by the legal science, and it can, equally, only be said to be existent or not existent. The point is that when the norm is said to be existent, this means that it is a valid norm (here "valid norm" is actually a pleonasm). The very existence of the norm in the legal system constitutes already its validity. The validity of a norm is therefore its specific existence.

25.1.2 Dichotomies in Kelsen's Theory

Hans Kelsen's philosophy of Law is essentially marked by its manifold dichotomies. When questions about the purposes of the legal sciences are at stake, another dichotomy then arises: the legal science is a descriptive science inasmuch as it doesn't make any prescription; but, at the same time, the objects of its descriptions are not statements, but prescriptions. Consequently, the normative science makes descriptions about prescriptions. Norberto Bobbio finely analyzes this issue:

> '*Normative' is in opposition (already in the Hauptprobleme) not to 'descriptive', but to 'explicative'; and, in parallel, 'descriptive' is in opposition (especially in the last works) not to 'normative', but to 'prescriptive'. Given that the doubles 'normative-explicative' and 'prescriptive-descriptive' do not superimpose themselves, there is no contradiction in affirming, as Kelsen does, that the legal science is at the same time descriptive and normative: descriptive in the sense that it does not prescribe, normative in the sense that the things being described are not facts, but norms, i.e., it is descriptive not about what*

exists, but about what ought to be. As Sollsätze, the propositions which characterize the legal science are distinguished in one hand from the Seinsätze belonging to social sciences (causal), and, on the other hand, from the Sollnormen of any normative system.[13]

What Bobbio tries to explain is that the normative science, despite the fact of dealing with norms – where the "normative" term comes from –, does not make use of a prescriptive discourse, that is, its descriptions don't have the aim of changing the behavior of others. They are statements capable of being evaluated or verified, and they are placed in the factual domain of the "is".

25.2 The Basic Norm as a Scientific Fiction Without Prescriptive Value

After the preceding considerations, one should be tended to imagine what the legal science really looks like. Kelsen sustains a hierarchical vision of the legal system (the object of the legal science), in the form of a triangle or a pyramid, where the base is formed by the particular norms created[14] by the Judge in the tribunals. Those norms must be based on more general norms, until we arrive at the Constitution of a Country, for example. And one could still regress and go up on the pyramid to attain the first Constitution of a Country. The role of a legal theory is to scientifically describe those norms as an object of study, as a system. The Legal Positivism (contrary to the Realism or Naturalism) focus on the claim that the norm and the Law are human constructions – they must be posited, enacted by someone – and the central notion is not the efficacy of a norm, or the moral value of a norm, but its validity, its existence in the legal system.

25.2.1 The Searching for Justification

A special question that arises in the justification of the normative system refers to what would make a unity from the multiple norms of a Country, for example. Well, the first response to that would probably mention the Constitution. The particular norms applied by the Judge (by his objective act of will) in the tribunals would be supported by the general norms created by the Legislator (by his objective act of will), and that's how we arrive at the Constitution. But here we have to face a problem: from where does the Constitution obtain its validity? From a prior

[13]BOBBIO, N. *Direito e Poder*. Editora Unesp, São Paulo, 2008, p. 58. This quotation was a personal translation of a Portuguese version of this book.

[14]Maybe it would be more correctly said that the Judge doesn't create any norm, he only applies the norms present on the system. But it won't be entirely wrong to say that particular or individual norms are actually created by the Judge.

Constitution, one could say. But from where does the first Constitution of a Country (that earlier here we have putted in the top of the pyramid) obtain its validity? Or, in other words, the person who enacted the first Constitution of a Country was authorized by whom to do so? How did it get legal legitimacy?

Here's the difficulty: every time we will put a prior legal document in this regression, the question about its legitimacy will be at stake. To solve this problem, Kelsen will say that what gives the foundations of the legal system cannot be a positive legal norm, a written document, because this enacted norm would forcedly have to be supported by a previous one, and the person enacting this norm would have to obtain the power and authority to create it from anyone coming from a higher degree. So, Kelsen will say, what legitimizes the creation of the first Constitution of a Country or, more generally, what legitimizes and gives the unity to a whole positive legal system, is in fact not a positive, but a fictive norm, called the Basic Norm.[15]

What we will advocate here is that this Basic Norm is not placed in the legal system, on the top of the pyramid, as many experts and critics of Kelsen suggest. We will defend that the Basic Norm is nothing more than a scientific fiction, a methodological tool. But, more precisely, our original approach will suggest that this regression in the seeking for the legitimation of the validity of the whole system must stop at the Constitution, and what we actually need after arriving there is not a new element, but a scientific tool used to conveniently consider this last object according to the positivist claims. It's only when we arrive at the top of the pyramid that we finally need the notion of Basic Norm, in order to rationally consider the first Constitution of a Country as the elementary component of the legal order.

The fact of clearly placing the Basic Norm in the scientific level will bring many clarifications to the kelsenian approach. First, the principle of purity will not be harmed by the presence of a fiction. Second, the theoretical aim of the Basic Norm will be preserved, since it's clear that the Basic Norm is not a norm to be respected by the individuals – it is not an effective norm. Lastly, the only problem that could remain, and the more difficult one, will be related to questions of the form: "But why do Kelsen call the Basic Norm a *norm*?". Our answer to this question makes reference to the fact that the fictive character of the Basic Norm concerns precisely its validity (and, as far as we remember, "validity" equals to "specific existence of a norm"), so it is the validity of this "norm" that makes reference a fiction. But what is fictional about the Basic Norm?

25.2.2 Understanding the Fiction

There is a difference between the fiction of the Basic Norm and the usual fictions present in Law (among the other norms of the legal system). The most common

[15]If we were to (wrongly) give some formulation to the Basic Norm, for didactic purposes let's say, it would sound like "We should do as the first Constitution of this Country tells us to do".

types of legal fiction will include contradictory elements in its formulation, leading to an apparently impossible situation, with the objective of attaining a specific goal. Let's see some examples:

> *In property law, a husband and wife could be treated as one person. In family law, a child's will is attributed to his guardian. In the interpretation of wills, one spouse may be deemed to have predeceased the other, even though that may not in fact have been the case. Under the attractive nuisance doctrine, a child who trespasses is treated as having been invited onto the defendant's land. In immigration law, an alien may be considered to be legally excluded from the United States even though he is physically within its borders. In civil forfeiture proceedings, property itself may be named as a party to litigation (. . .).*[16]

Clearly, the Basic Norm doesn't share the same characteristics of the so-called legal fictions, that is to say the norms including contradictory elements in their formulation. So, where is the fiction concerning the Basic Norm? How can we recognize it as a fiction?

To create the notion of Basic Norm as a fiction (a conception which is present only in the last writings of Kelsen – before the Basic Norm was seen as an hypothesis[17]), Kelsen searches for the foundations of his conception of fiction in the work "The Philosophy of As-If",[18] from the German philosopher Hans Vaihinger (1852–1933). To Kelsen, the fictional element concerns the *act* of conceiving the foundations for the legal system: we make "as if" there were a higher norm above the first Constitution of a State, in order to stop the searching for legitimacy and to give to the system a unity.

So, in the case of the Basic Norm, the fictive element cannot be placed in the formulation of the norm. To formulate the Basic Norm in a normative proposition is to submit it to a higher justification, is to presuppose that there is a higher power above it, to legitimize its creation. So, it's not the case that its formulation will give rise to contradictions, but the fact of being formulated will abort its very function. Once formulated, the Basic Norm is legally enacted and enters in the circular searching for its own legitimacy. Kelsen himself explains this difference, showing that the fact of accepting the necessity of a fictive Basic Norm does not imply accepting the presence of fictions in Law:

> *According to Vaihinger, a fiction is a cognitive device used when one is unnabled to attain one's cognitive goal with the material at hand (1935:13). The cognitive goal of the Basic Norm is to ground the validity of the norms forming a positive moral or legal order, that*

[16] Student Author. Harvard Law Review, Vol. 115, No. 8, pp. 2233–2234, Jun. 1918.

[17] In his previous writings (KELSEN, H. "Théorie Pure du Droit", 2e traduction par Ch. EISENMANN, Dalloz, Paris 1962), Kelsen tended to see the Basic Norm not as a fiction, but as a hypothesis. This approach was abandoned because Kelsen got aware from the fact that the "existence" of the Basic Norm would never be able to be verified. Contrary to hypotheses which are constructed with the very aim to be lately confirmed or falsified, the Basic Norm is presupposed with the consciousness of its impossibility of ever being verified.

[18] Vaihinger, H. *The Philosophy of "As-If": a system of the theoretical, practical and religious fictions of mankind*. London, Routledge & K. Paul, 1965 [1911].

is, to interpret the subjective meaning of the norm-positing acts as their objective meaning (i.e. as valid norms) and to interpret the relevant acts as norm-positing acts. This goal can be attained only by means of a fiction. It should be noted that the Basic Norm is not a hypothesis in the sense of Vaihinger's philosophy of As-If – as I myself have sometimes characterized it – but a fiction. A fiction differs from a hypothesis in that it is accompanied – or ought to be accompanied – by the awareness that reality does not agree with it. (Kelsen 2011[1979], p. 256)

25.2.3 Where Is the Basic Norm?

If we cannot formulate the Basic Norm in the terms of a prescription, it clearly means that it is not a norm, a valid norm or an efficient norm. We cannot obey or disrespect the Basic Norm. Its purpose is not that of controlling our behavior. Well, that is the key to understand the notion of Basic Norm in Kelsen's theory. We claim that the searching for the justification stops at the First Constitution of a State. Whatever comes after concerns no longer the Law, but only the legal science. And the Basic Norm has no interest to the lawyer or the judge, to the defendant or the complainant. It is a matter of scientific understanding of the system, in the terms that the legal system could not be rationally described without the notion of the Basic Norm. The Basic Norm appears in the act of the jurist when he is confronted to the problem of the formal justification of the whole system of legal norms. The jurist need to presuppose the existence of a fiction of the Basic Norm, otherwise the legal system would be nothing but a chaotic heap of norms.

Our point is to clear up the fact that the Basic Norm finds its fictive element in the act of doing "as if" it existed. To conceive the Basic Norm consists in an act that involves the presumption of its existence, even though we know that the Basic Norm cannot exist. The fictiveness of this "norm" opposes anyone to identify it in any level: the Basic Norm is not a positive norm, but neither a formulated principle found in legal science. The point is that nobody is able to formulate, to construct such a thing as the Basic Norm. If it is enacted, we enter in the vicious circle and we need a new element to justify its existence as a norm. If it is formulated as a scientifically principle, it loses its purpose, which is to end the regression for justification and serve as the base for the legal system. What we defend is that the key to understand the needing of a Basic Norm is to see it as a part of the process of justifying the system, as a scientific attitude in front of the legal structure. To put it in another way: we only can understand the legal system as a rational ensemble of norms from the moment when we presuppose the fiction of a Basic Norm as being the starting point of the normative creation. The presupposition of the Basic norm is the condition *sine qua non* for the legal system to be a valid one (as a whole).

That's why do we have to call the Basic Norm a norm. The fiction about the Basic Norm relies exactly in its validity. We have to make "as if" there was a Basic Norm, because we know that it cannot be valid, it cannot legally exist. But we make

"as if" it could give the legislator the power to enact the first Constitution, to turn the Constitution into a binding document.[19]

This approach can answer a famous critic to the notion of Basic Norm: that the whole positivist legal system will finally be built over a fiction. We defend here that the whole dynamic structure of legal power related to the normative creation is in fact founded on a fiction, but not the legal normative system, namely, the set of legal norms themself. The searching for the justification of the validity of the norm can be retraced until the first Constitution of the Country, and the process stops there. Materially, there is no legal element before the first Constitution – the Basic Norm has no legal validity. We need the concept of Basic Norm only in a scientific level, in order to give the system its unity and its legitimacy as a whole.

So, to answer our question from this section, the Basic Norm is not a part of the legal system, but it is neither a part of the legal science. It is a presupposition from the jurist, prior to any consideration about a specific legal system. Without conceiving the notion of Basic norm, no legal system can be rationally analyzed, studied or interpreted by a science. Without the fiction of a Basic Norm, the legal system will be just a multiple set of norms without inner cohesion, without beginning or end. The Basic Norm, according to our approach, should be seen as an inherent methodological procedure, prior to any consideration of a positive legal system.

25.3 Conclusion

The fiction of the Basic Norm has to be clearly separated from the cases of legal fictions that we know since Roman law. The Basic Norm has no legal validity, has no regulating power in relation to our acts, it can't even be formulated. In order to dissolve the confusion saying that Kelsen loses track in inserting a non-positive norm in its legal theory, we must understand that the Basic Norm is nothing more than a scientific methodological device, to be used by the scientist. It has no direct relation with the ordinary norms of a system, it only works as a presupposition from the part of the jurist when it's question of considering a specific order as a whole valid and legitimate legal order. It is a tool that allows the scientist to approach the legal system and, if we must make a metaphor here, we could use the beautiful image of Kant to say that the legal system can only be rationally regarded through the lenses of the Basic Norm.

[19]We must remember that the first aim in using a fiction in science is always its practical utility. Loewenberg (Review, The Journal of Philosophy, Psychology and Scientific Methods, Vol. 9, No. 26 (Dec. 19, 1912), p. 717) explains clearly the notion of fiction according to Vaihinger's theory: "Fictions, in Vaihinger's usage, are not identical with figments, such as centaur or fairy, nor are they hypotheses capable of verification. They are **deliberate devices (*Kunstgriffe*)** on the part of thought for the practical purpose of successful orientation in and perfect control over the environment. Theoretically they are absolutely valueless. Applied with a knowledge of their fictitious character, they will lead to the intended practical results."

The fictive character of the norm is explained by Kelsen's fascination with the work of Hans Vaihinger. Vaihinger's "Philosophy of as-if" is perfectly adapted to what Kelsen wanted for the Basic Norm: to show that the use of a fiction in the frame of a science is perfectly viable in order to attain an objective that could not be attained otherwise: give to the legal system the unity necessary to the legal analysis. To completely understand the necessity of the Basic Norm is to accept that it has to be completely separated from the legal system; it's to not being misled by the "norm" in Basic Norm.

When Kelsen adopted Vaihinger's "philosophy of as-if", he was certainly attracted by this notion of "*making* as-if" something were the case, and what we try to clear up in this text is the emphasis on the "*act"* of doing as if the Basic Norm existed. The Basic Norm is something that we *use*, it is a scientific *tool*. It is not a legal element and, if we stay strict, not even a methodological element, in the sense that it doesn't really make part of the legal science, but it must be a part, and it is an essential element, of the scientist's *approach* towards its objects, the legal norms. Without the presupposition of a Basic Norm by the jurist, or the Judge, or the Legislator, no theory can be founded, no science can exist.

References

Bobbio, N.: Teoría General del Derecho. Editorial Temis, Santa Fé de Bogotá (1997)
Bobbio, N.: Direito e Poder. Editora Unesp, São Paulo (2008)
Hart, H.: Visita a Kelsen. In: Lua Nova. No. 64. São Paulo Jan./Apr. 2005 (1963)
Kelsen, H.: Théorie Pure du Droit. Translation: EISENMANN, C. Dalloz, Paris (1962)
Kelsen, H.: General Theory of Norms. Oxford University Press, New York 2011 (1979)
Loewenberg, J.: Reviewed work(s): Die Philosophie des Als Ob. System der Theoretischen, Praktischen und Religiösen Fiktionen der Menschheit auf Grund Eines Idealistischen Positivismus. Mit einem Anhang über Kant und Nietzsche by H. Vaihinger. J. Philos. Psychol. Scientific Methods **9**(26) (Dec. 19, 1912). Published by: Journal of Philosophy, Inc. Stable URL: http://www.jstor.org/stable/2013048 (1912)
Spaak, T.: Kelsen and Hart on the normativity of law. In: Wahlgren, P. (ed.) Perspectives on Jurisprudence: Essays in Honour of Jes Bjarup (2005)
Student Author.: Harvard Law Rev. **115**(8) 2233–2234, June (1918)
Vaihinger, H.: The Philosophy of 'As-If': A System of the Theoretical, Practical and Religious Fictions of Mankind. Routledge & K. Paul, London 1965 (1911)
van Roermund, B.: Authority and authorization. In: Law and Philosophy. Springer, Vol. 19, No. 2, pp. 201–222, Mar. (2000)

Author Index

J. Redmond et al. (eds.), *Epistemology, Knowledge and the Impact of Interaction*,
Logic, Epistemology, and the Unity of Science 38, DOI 10.1007/978-3-319-26506-3

Subject Index

J. Redmond et al. (eds.), *Epistemology, Knowledge and the Impact of Interaction*,
Logic, Epistemology, and the Unity of Science 38, DOI 10.1007/978-3-319-26506-3