HANDBOOK OF PHILOSOPHICAL LOGIC, 2ND EDITION

Editors:
D.M. Gabbay and F. Guenthner

VOLUME 14

Springer

HANDBOOK OF PHILOSOPHICAL LOGIC
2ND EDITION

VOLUME 14

HANDBOOK OF PHILOSOPHICAL LOGIC

2nd Edition

Volume 14

edited by D.M. Gabbay and F. Guenthner

Volume 1 – ISBN 0-7923-7018-X
Volume 2 – ISBN 0-7923-7126-7
Volume 3 – ISBN 0-7923-7160-7
Volume 4 – ISBN 1-4020-0139-8
Volume 5 – ISBN 1-4020-0235-1
Volume 6 – ISBN 1-4020-0583-0
Volume 7 – ISBN 1-4020-0599-7
Volume 8 – ISBN 1-4020-0665-9
Volume 9 – ISBN 1-4020-0699-3
Volume 10 – ISBN 1-4020-1644-1
Volume 11 – ISBN 1-4020-1966-1
Volume 12 – ISBN 1-4020-3091-6
Volume 13 – ISBN 978-1-4020-3520-3

HANDBOOK OF PHILOSOPHICAL LOGIC

2nd EDITION

VOLUME 14

Edited by

D.M. GABBAY
King's College, London, U.K.

and

F. GUENTHNER
Centrum für Informations- und Sprachverarbeitung,
Ludwig-Maximilians-Universität München, Germany

Springer

A C.I.P. Catalogue record for this book is available from the Library of Congress.

ISBN 978-1-4020-6323-7 (HB)
ISBN 978-1-4020-6324-4 (e-book)

Published by Springer,
P.O. Box 17, 3300 AA Dordrecht, The Netherlands.

www.springer.com

Printed on acid-free paper

CONTENTS

PREFACE TO THE SECOND EDITION

It is with great pleasure that we are presenting to the community the second edition of this extraordinary handbook. It has been over 15 years since the publication of the first edition and there have been great changes in the landscape of philosophical logic since then.

The first edition has proved invaluable to generations of students and researchers in formal philosophy and language, as well as to consumers of logic in many applied areas. The main logic article in the Encyclopaedia Britannica 1999 has described the first edition as 'the best starting point for exploring any of the topics in logic'. We are confident that the second edition will prove to be just as good!

The first edition was the second handbook published for the logic community. It followed the North Holland one volume *Handbook of Mathematical Logic*, published in 1977, edited by the late Jon Barwise. The four volume *Handbook of Philosophical Logic*, published 1983–1989 came at a fortunate temporal junction at the evolution of logic. This was the time when logic was gaining ground in computer science and artificial intelligence circles.

These areas were under increasing commercial pressure to provide devices which help and/or replace the human in his daily activity. This pressure required the use of logic in the modelling of human activity and organisation on the one hand and to provide the theoretical basis for the computer program constructs on the other. The result was that the *Handbook of Philosophical Logic*, which covered most of the areas needed from logic for these active communities, became their bible.

The increased demand for philosophical logic from computer science and artificial intelligence and computational linguistics accelerated the development of the subject directly and indirectly. It directly pushed research forward, stimulated by the needs of applications. New logic areas became established and old areas were enriched and expanded. At the same time, it socially provided employment for generations of logicians residing in computer science, linguistics and electrical engineering departments which of course helped keep the logic community thriving. In addition to that, it so happens (perhaps not by accident) that many of the Handbook contributors became active in these application areas and took their place as time passed on, among the most famous leading figures of applied philosophical logic of our times. Today we have a handbook with a most extraordinary collection of famous people as authors!

The table below will give our readers an idea of the landscape of logic and its relation to computer science and formal language and artificial intelligence. It shows that the first edition is very close to the mark of what was needed. Two topics were not included in the first edition, even though

they were extensively discussed by all authors in a 3-day Handbook meeting. These are:

- a chapter on non-monotonic logic
- a chapter on combinatory logic and λ-calculus

We felt at the time (1979) that non-monotonic logic was not ready for a chapter yet and that combinatory logic and λ-calculus was too far removed.[1] Non-monotonic logic is now a very major area of philosophical logic, alongside default logics, labelled deductive systems, fibring logics, multi-dimensional, multimodal and substructural logics. Intensive re-examinations of fragments of classical logic have produced fresh insights, including at time decision procedures and equivalence with non-classical systems.

Perhaps the most impressive achievement of philosophical logic as arising in the past decade has been the effective negotiation of research partnerships with fallacy theory, informal logic and argumentation theory, attested to by the Amsterdam Conference in Logic and Argumentation in 1995, and the two Bonn Conferences in Practical Reasoning in 1996 and 1997.

These subjects are becoming more and more useful in agent theory and intelligent and reactive databases.

Finally, fifteen years after the start of the Handbook project, I would like to take this opportunity to put forward my current views about logic in computer science, computational linguistics and artificial intelligence. In the early 1980s the perception of the role of logic in computer science was that of a specification and reasoning tool and that of a basis for possibly neat computer languages. The computer scientist was manipulating data structures and the use of logic was one of his options.

My own view at the time was that there was an opportunity for logic to play a key role in computer science and to exchange benefits with this rich and important application area and thus enhance its own evolution. The relationship between logic and computer science was perceived as very much like the relationship of applied mathematics to physics and engineering. Applied mathematics evolves through its use as an essential tool, and so we hoped for logic. Today my view has changed. As computer science and artificial intelligence deal more and more with distributed and interactive systems, processes, concurrency, agents, causes, transitions, communication and control (to name a few), the researcher in this area is having more and more in common with the traditional philosopher who has been analysing

[1] I am really sorry, in hindsight, about the omission of the non-monotonic logic chapter. I wonder how the subject would have developed, if the AI research community had had a theoretical model, in the form of a chapter, to look at. Perhaps the area would have developed in a more streamlined way!

such questions for centuries (unrestricted by the capabilities of any hardware).

The principles governing the interaction of several processes, for example, are abstract an similar to principles governing the cooperation of two large organisation. A detailed rule based effective but rigid bureaucracy is very much similar to a complex computer program handling and manipulating data. My guess is that the principles underlying one are very much the same as those underlying the other.

I believe the day is not far away in the future when the computer scientist will wake up one morning with the realisation that he is actually a kind of formal philosopher!

The projected number of volumes for this Handbook is about 18. The subject has evolved and its areas have become interrelated to such an extent that it no longer makes sense to dedicate volumes to topics. However, the volumes do follow some natural groupings of chapters.

I would like to thank our authors and readers for their contributions and their commitment in making this Handbook a success. Thanks also to our publication administrator Mrs J. Spurr for her usual dedication and excellence and to Springer for their continuing support for the Handbook.

Dov Gabbay
King's College London

Logic	**IT**			
	Natural language processing	**Program control specification, verification, concurrency**	**Artificial intelligence**	**Logic programming**
Temporal logic	Expressive power of tense operators. Temporal indices. Separation of past from future	Expressive power for recurrent events. Specification of temporal control. Decision problems. Model checking.	Planning. Time dependent data. Event calculus. Persistence through time—the Frame Problem. Temporal query language. temporal transactions.	Extension of Horn clause with time capability. Event calculus. Temporal logic programming.
Modal logic. Multi-modal logics	generalised quantifiers	Action logic	Belief revision. Inferential databases	Negation by failure and modality
Algorithmic proof	Discourse representation. Direct computation on linguistic input	New logics. Generic theorem provers	General theory of reasoning. Non-monotonic systems	Procedural approach to logic
Non-monotonic reasoning	Resolving ambiguities. Machine translation. Document classification. Relevance theory	Loop checking. Non-monotonic decisions about loops. Faults in systems.	Intrinsic logical discipline for AI. Evolving and communicating databases	Negation by failure. Deductive databases
Probabilistic and fuzzy logic	logical analysis of language	Real time systems	Expert systems. Machine learning	Semantics for logic programs
Intuitionistic logic	Quantifiers in logic	Constructive reasoning and proof theory about specification design	Intuitionistic logic is a better logical basis than classical logic	Horn clause logic is really intuitionistic. Extension of logic programming languages
Set theory, higher-order logic, λ-calculus, types	Montague semantics. Situation semantics	Non-well-founded sets	Hereditary finite predicates	λ-calculus extension to logic programs

Imperative vs. declarative languages	**Database theory**	**Complexity theory**	**Agent theory**	**Special comments: A look to the future**
Temporal logic as a declarative programming language. The changing past in databases. The imperative future	Temporal databases and temporal transactions	Complexity questions of decision procedures of the logics involved	An essential component	Temporal systems are becoming more and more sophisticated and extensively applied
Dynamic logic	Database updates and action logic	Ditto	Possible actions	Multimodal logics are on the rise. Quantification and context becoming very active
Types. Term rewrite systems. Abstract interpretation	Abduction, relevance	Ditto	Agent's implementation rely on proof theory.	
	Inferential databases. Non-monotonic coding of databases	Ditto	Agent's reasoning is non-monotonic	A major area now. Important for formalising practical reasoning
	Fuzzy and probabilistic data	Ditto	Connection with decision theory	Major area now
Semantics for programming languages. Martin-Löf theories	Database transactions. Inductive learning	Ditto	Agents constructive reasoning	Still a major central alternative to classical logic
Semantics for programming languages. Abstract interpretation. Domain recursion theory.		Ditto		More central than ever!

Classical logic. Classical fragments	Basic background language	Program synthesis	A basic tool	
Labelled deductive systems	Extremely useful in modelling		A unifying framework. Context theory.	Annotated logic programs
Resource and substructural logics	Lambek calculus		Truth maintenance systems	
Fibring and combining logics	Dynamic syntax	Modules. Combining languages	Logics of space and time	Combining features
Fallacy theory				
Logical Dynamics	Widely applied here			
Argumentation theory games		Game semantics gaining ground		
Object level/ metalevel			Extensively used in AI	
Mechanisms: Abduction, default relevance			ditto	
Connection with neural nets				
Time-action-revision models			ditto	

	Relational databases	Logical complexity classes	The workhorse of logic	The study of fragments is very active and promising.
	Labelling allows for context and control.		Essential tool.	The new unifying framework for logics
Linear logic			Agents have limited resources	
	Linked databases. Reactive databases		Agents are built up of various fibred mechanisms	The notion of self-fibring allows for self-reference
				Fallacies are really valid modes of reasoning in the right context.
			Potentially applicable	A dynamic view of logic
				On the rise in all areas of applied logic. Promises a great future
			Important feature of agents	Always central in all areas
			Very important for agents	Becoming part of the notion of a logic
				Of great importance to the future. Just starting
			A new theory of logical agent	A new kind of model

WALTER CARNIELLI, MARCELO E. CONIGLIO AND
JOÃO MARCOS

LOGICS OF FORMAL INCONSISTENCY

1 INTRODUCTION

1.1 Contradictoriness and inconsistency, consistency and non-contradictoriness

In traditional logic, contradictoriness (the presence of contradictions in a theory or in a body of knowledge) and triviality (the fact that such a theory entails all possible consequences) are assumed inseparable, granted that negation is available. This is an effect of an ordinary logical feature known as 'explosiveness': According to it, from a contradiction 'α and $\neg\alpha$' everything is derivable. Indeed, classical logic (and many other logics) equate 'consistency' with 'freedom from contradictions'. Such logics forcibly fail to distinguish, thus, between contradictoriness and other forms of inconsistency. Paraconsistent logics are precisely the logics for which this assumption is challenged, by the rejection of the classical 'consistency presupposition'. The *Logics of Formal Inconsistency*, **LFI**s, object of this chapter, are the paraconsistent logics that neatly balance the equation:

$$\text{CONTRADICTIONS} + \text{CONSISTENCY} = \text{TRIVIALITY}$$

The **LFI**s have a remarkable way of reintroducing consistency into the non-classical picture: They internalize the very notions of consistency and inconsistency at the object-language level. The result of that strategy is the design of very expressive logical systems, whose fundamental feature is the ability to recover all consistent reasoning right on demand, while still allowing for some inconsistency to linger, otherwise.

Paraconsistency is the study of contradictory yet non-trivial theories.[1] The significance of paraconsistency as a philosophical program which dares to go beyond consistency lies in the possibilities (formal, epistemological and mathematical) to take profit from the distinctions and contrasts between asserting opposites (either in a formal or in a natural language) and ensuring non-triviality (in a theory, formal or not). A previous entry [Priest, 2002] in this Handbook was dedicated to paraconsistent logics. Although partaking in the same basic views on paraconsistency, our approach is oriented towards investigating and exhibiting the features of an ample and very expressive class of paraconsistent logics — the above mentioned **LFI**s.

[1] Paraconsistency has the meaning of 'besides, beyond consistency', just as paradox means 'besides, beyond opinion' and 'paraphrase' means 'to phrase in other words'.

D.M. Gabbay and F. Guenthner (eds.), Handbook of Philosophical Logic, 2nd Edition, Volume 14, 1–93.

Moreover, our chapter starts from clear-cut abstract definitions of the terms involved (triviality, consistency, paraconsistency, etc.) and analyzes both proof-theoretical and model-theoretical aspects of **LFI**s, insisting on their special interest and hinting about their near ubiquity in the paraconsistent realm.

Once inconsistency is locally allowed, the chief value of a useful logical system (understood as a derivability formalism reflecting some given theoretical or pragmatical constraints) turns out to be its capability of doing what it is supposed to do, namely, to set acceptable inferences apart from unacceptable ones. The least one would ask for is, thus, that the system *does* separate propositions (into two non-empty classes, the derivable ones and the non-derivable) or, in other words, that it be non-trivial. Therefore, the most fundamental guiding criterion for choosing theories and systems worthy of investigation, as suggested by [Jaśkowski, 1948], [Nelson, 1959] and [da Costa, 1959], and extended in [Marcos, 2005c], should indeed be their abstract character of non-triviality, rather than the mere absence of contradictions.

The big challenge for paraconsistentists is to avoid allowing contradictory theories to explode and derive anything else (as they do in classical logic) and *still* to reserve resources for designing a respectful logic. For that purpose they must weaken their logical machinery by abandoning explosion in order to be able to draw reasonable conclusions from those theories, and *yet* come up with a legitimate logical system. A current trend in logic has been that of internalizing metatheoretical notions and devices at the object-language level, in order to build ever more expressive logical systems, as in the case of labeled deductive systems, hybrid logics, or the logics of provability. The **LFI**s constitute exactly the class of paraconsistent logics which can internalize the metatheoretical notions of consistency and inconsistency. As a consequence, despite constituting fragments of consistent logics, the **LFI**s can canonically be used to faithfully encode all consistent inferences. We will in this chapter present and discuss these logics, illustrating their uses, properties and representations.

Most of the material for the chapter is based on the article [Carnielli and Marcos, 2002], which founds the formal distinctions between contradictoriness, inconsistency and triviality, which we here utilize. In some cases we correct here the definitions and proofs presented there. Another central reference is the book [Marcos, 2005], where most of the examples and proposals hereby defended may be found, in extended form. The **LFI**s, central topic of the present chapter, are carefully introduced in Subsection 3.1. All necessary concepts and definitions showing how we approach the property of explosion and how this reflects on the principles of logic will be found in Section 2. Subsection 1.2 serves as vestibular to the more technical sections that follow.

The main **LFI**s are presented in Sections 3 and 4. One of their primary

subclasses, the **C**-systems, is introduced as containing those **LFI**s in which consistency may be expressed as a single formula of the object language. Moreover, the **dC**-systems are introduced as those **C**-systems in which this same formula may be explicitly expressed in terms of other more usual connectives (see Definition 32). In Section 3 we study in detail a fundamental example of **LFI**, the logic **mbC**, where consistency is rendered expressible by means of a specific new primitive connective. This logic is compared to the stronger logic C_1 (cf. [da Costa, 1963] and [da Costa, 1974]), a logic of the early paraconsistent vintage. We provide Hilbert-style axiomatizations, as well as bivaluation semantics and adequate tableau systems for **mbC** and C_1. Additionally, adequate possible-translations semantics are proposed for **mbC**.

LFIs are typically based on previously given consistent logics. The fundamental feature enjoyed by classically-based **LFI**s of being able to recover classical reasoning (despite constituting themselves deductive fragments of classical logic) is explained in Subsection 3.6.

In Section 4 we extend the logic **mbC** by adding further axioms which permit us to talk about inconsistency and consistency in more symmetric guises inside the logic. A brief study of the thereby obtained logics follows, extending the results obtained in Section 3.

Section 5 explores additional topics on **LFI**s. In Subsection 5.1 some fundamental **dC**-systems are studied. Particular cases of **dC**-systems are da Costa's logics C_n, $1 \leq n < \omega$, Jaśkowski's logic **D2**, and all usual normal modal logics (under convenient formulations). Conveniently extending the previously obtained **LFI**s it is possible to introduce a large family of such logics by controlling the propagation of consistency (cf. Subsection 5.2). This procedure adds flexibility to the game, allowing one to propose tailor-suited **LFI**s; we illustrate the case by defining literally thousands of logics, including an interesting class of maximal logics in Subsection 5.3. We end this subsection by a brief note on the possibilities of algebraizing **LFI**s, in general, concluding a series of similar notes and results to be found along the paper, dedicated especially to the difficulties surrounding the so-called replacement property, the metatheoretical result that guarantees equivalent formulas to be logically indistinguishable.

Section 6 examines some perspectives on the research about Logics of Formal Inconsistency. The chapter ends by a list of axioms and systems given in Section 7.

It goes without saying that the route we will follow in this chapter corresponds not only to our preferences on how to deal with paraconsistency, but it brings also a personal choice of topics we consider to be of special philosophical and mathematical relevance.

1.2 The import of the Logics of Formal Inconsistency

Should the presence of contradictions make it impossible to derive anything sensible from a theory or a logic where such contradictions appear, as the classical logician would maintain? Or are there maybe situations in which contradictions are at least temporarily admissible, if only their wild behavior can somehow be controlled? The theoretical and practical relevance of such questions shows paraconsistency to be a bold programme in the foundations of formal sciences. As time goes by, the problems and methods of formal logic, traditionally connected to mathematics and philosophy, can more and more be seen to affect and influence several other areas of knowledge, such as computer science, information systems, formal philosophy, theoretical linguistics, and so forth. In such areas, certainly more than in mathematics, contradictions are presumably unavoidable: If contradictory theories appear only by mistake, or are due to some kind of resource-boundedness on computers, or depend on an altered state of reality, contradictions can hardly be prevented from at least being taken into consideration, as they often show up as gatecrashers. The pragmatic point thus is not whether contradictory theories *exist*, but *how to deal with them*.

Regardless of the disputable status of contradictory theories, it is hard to deny that they are, in many cases, quite *informative*, it being desirable to establish *well-reasoned* judgements even when contradictions are present. Consider, for instance, the following situation (adapted from [Carnielli and Marcos, 2001a]) in which you ask a yes-no question to two people: 'Does Jeca Tatu live in São Paulo?' Exactly one of the three following distinct scenarios is possible: They might both say 'yes', they might both say 'no', or else one of them might say 'yes' while the other says 'no'. Now, it happens that in no situation you can be sure whether Jeca Tatu lives in São Paulo or not (unless you trust one of the interviewees more than the other), but only in the last scenario, where a contradiction appears, you are sure to have received wrong information from one of your sources.

A challenge to any study on paraconsistency is to oppugn the tacit assumption that contradictory theories necessarily contain false sentences. Thus, if we can build models of structures in which some (but not all) contradictory sentences are simultaneously true, we will have the possibility of maintaining contradictory sentences inside a given theory and still be able, in principle, to perform reasonable inferences from that theory. The problem will not be that of *validating falsities*, but rather of *extending our notion of truth* (an idea further explored, for instance, in [Bueno, 1999]).

In the first half of the last century, some authors, including Łukasiewicz and Vasiliev, proposed a new approach to the idea of non-contradiction, offering interpretations to formal systems in which contradictions could make sense. Between the 1940s and the 60s the first systems of paraconsistent logic appeared (cf. [Jaśkowski, 1948], [Nelson, 1959], and [da Costa,

1963]). For historical notes on paraconsistency we suggest [Arruda, 1980], [D'Ottaviano, 1990], [da Costa and Marconi, 1989], the references mentioned in part 1 of [Priest *et al.*, 1989] and in section 3 of [Priest, 2002], as well as the book [Bobenrieth-Miserda, 1996] and the prolegomena to [Marcos, 2005].

Probably around the 40s, time was ripe for thinking about the role of negation in different terms: The falsificationism of K. Popper (cf. [Popper, 1959]) supported the idea (and stressed its role in the philosophy of science) that falsifying a proposition, as an epistemological step towards refuting it, is not the same as assuming the sentence to be false. This apparently led Popper to think about a paraconsistent-like logic dual to intuitionism in his [Popper, 1948], later to be rejected as somehow too weak as to be useful (cf. [Popper, 1989]). But it should be remarked that Popper never dismissed this kind of approach as nonsensical. His disciple D. Miller in [Miller, 2000] in fact argues that the logic for dealing with unfalsifiedness should be paraconsistent.[2] Another recent proposal by Y. Shramko also defends the paraconsistent character of falsificationism (cf. [Shramko, 2005]).

When proposing his first paraconsistent logics (cf. [da Costa, 1963]) da Costa's intuition was that the 'consistency' (which he dubbed 'good behavior') of a given formula would not only be a sufficient requisite to guarantee its explosive character, but that it could also be represented as an ordinary formula of the underlying language. For his initial logic, C_1, he chose to represent the consistency of a formula α by the formula $\neg(\alpha \wedge \neg\alpha)$, and referred to this last formula as a realization of the 'Principle of Non-Contradiction'.

In the present approach, as in [Carnielli and Marcos, 2002], we introduce consistency as a *primitive notion* of our logics: The Logics of Formal Inconsistency, **LFI**s, are paraconsistent logics that internalize the notions of consistency and inconsistency at the object-language level. In this chapter we will also study some significative subclasses of **LFI**s, the **C**-systems and **dC**-systems based on classical logic (and da Costa's logics C_n will be shown to constitute but particular samples from the latter subclass).

It is worth noting that, in general, paraconsistent logics do not validate contradictions nor, equivalently, invalidate the 'Principle of Non-Contradiction', in our reading of it (cf. the principle (1) in Subsection 2.1). Most paraconsistent logics, in fact, are proper fragments of (some version of) classical logic, and thus they cannot be *contradictory*.

Clearly, the concept of paraconsistency is related to the properties of a negation inside a given logic. In that respect, arguments can be found in the literature to the effect that 'negations' of paraconsistent logics would not be proper negation operators (cf. [Slater, 1995] and [Béziau, 2002a]). Béziau's argument amounts to a request for the definition of some mini-

[2]Indeed, Miller even proposes that the logic C_1 of da Costa's hierarchy could be used as a logic of falsification.

mal 'positive properties' in order to characterize paraconsistent negation as constituting a real *negation* operator, instead of something else. Slater argues for the *inexistence* of paraconsistent logics, given that their negation operator is not a 'contradictory-forming functor', but just a 'subcontrary-forming one', revisiting and extending an earlier argument from [Priest and Routley, 1989]. A reply to the latter kind of criticism is that it is as convincing as arguing that a 'line' in hyperbolic geometry is not a real line, since, through a given point not on the line, the 'parallel-forming functor' does not define a unique line.[3] In any case, this is not the only possible counter-objection, and the development of paraconsistent logic is not deterred by this discussion. Investigations about the general properties of paraconsistent negations include [Avron, 2002], [Béziau, 1994] and [Lenzen, 1998], among others. Those studies are surveyed in [Marcos, 2005c], where also a minimal set of 'negative properties' for negation is advanced as a new starting point for a unifying study of negation.

2 WHY'S AND HOW'S: CONCEPTS AND DEFINITIONS

2.1 The principles of logic revisited

Our presentation in what follows is situated at the level of a general theory of consequence relations. Let $\wp(X)$ denote the powerset of a set X. As usual, given a set For of formulas, we say that $\Vdash \subseteq \wp(For) \times For$ defines a (single-conclusion) *S-consequence relation* over For (where S stands for *standard*) if the following clauses hold, for any choice of formulas α and β, and of subsets Γ and Δ of For (formulas and commas at the left-hand side of $\Vdash$ denote, as usual, sets and unions of sets of formulas):

(Con1) $\alpha \in \Gamma$ implies $\Gamma \Vdash \alpha$ (reflexivity)
(Con2) ($\Delta \Vdash \alpha$ and $\Delta \subseteq \Gamma$) implies $\Gamma \Vdash \alpha$ (monotonicity)
(Con3) ($\Delta \Vdash \alpha$ and $\Gamma, \alpha \Vdash \beta$) implies $\Delta, \Gamma \Vdash \beta$ (cut)

So, an *S-logic* $\mathbf{L}$ will here be defined simply as a structure of the form $\langle For, \Vdash \rangle$, containing a set of formulas For and an S-consequence relation $\Vdash$ defined over this set. An additional useful property of a logic is *compactness*, defined as:

(Con4) $\Gamma \Vdash \alpha$ implies $\Gamma^{\mathsf{fin}} \Vdash \alpha$, for some finite $\Gamma^{\mathsf{fin}} \subseteq \Gamma$ (compactness)

We will assume that the language of every logic $\mathbf{L}$ is defined over a propositional signature $\Sigma = \{\Sigma_n\}_{n \in \omega}$, where Σ_n is the set of connectives of arity n. We will also assume that $\mathcal{P} = \{p_n : n \in \omega\}$ is the set of propositional

[3] In hyperbolic geometry the following property, known as the Hyperbolic Postulate, holds good: For every line l and point p not on l, there exist at least two distinct lines parallel to l that pass through p.

variables (or atomic formulas) from which we freely generate the algebra For of formulas using Σ. Along most of the present paper, the least we will suppose on a *logic* is that its consequence relation satisfies the clauses defining an S-consequence.

Another usual property of a logic is *structurality*. Let ε be an endomorphism in For, that is, ε is the unique homomorphic extension of a mapping from $\mathcal{P}$ into For. A logic is structural if its consequence relation preserves endomorphisms:

(Con5) $\Gamma \Vdash \alpha$ implies $\varepsilon(\Gamma) \Vdash \varepsilon(\alpha)$ (structurality)

In syntactical terms, structurality corresponds to the rule of uniform substitution or, alternatively, to the use of schematic axioms and rules.

Any set $\Gamma \subseteq For$ is here called a *theory* of $\mathbf{L}$. A theory Γ is said to be *proper* if $\Gamma \neq For$, and a theory Γ is said to be *closed* if it contains all of its consequences, that is, for a closed theory Γ we have $\Gamma \Vdash \alpha$ iff $\alpha \in \Gamma$, for every formula α. If $\Gamma \Vdash \alpha$ for all Γ, we will say that α is a *thesis* (of $\mathbf{L}$).

Unless explicitly stated to the contrary, we will from now on be working with some fixed arbitrary logic $\mathbf{L} = \langle For, \Vdash \rangle$ where For is written in a signature containing a unary 'negation' connective $\neg$ and $\Vdash$ satisfies (Con1)–(Con3) and (Con5).

Let Γ be a theory of $\mathbf{L}$. We say that Γ is *contradictory with respect to* $\neg$, or simply *contradictory*, if it satisfies:

$$\exists \alpha(\Gamma \Vdash \alpha \text{ and } \Gamma \Vdash \neg\alpha)$$

(The formal framework to deal with this kind of metaproperties can be found in [Coniglio and Carnielli, 2002].) For any such formula α we may also say that Γ is *α-contradictory*.

A theory Γ is said to be *trivial* if it satisfies:

$$\forall \alpha(\Gamma \Vdash \alpha)$$

Of course the theory For is trivial, given (Con1). We can immediately conclude that contradictoriness is a necessary (but, in general, not a sufficient) condition for triviality in a given theory, since a trivial theory derives everything.

A theory Γ is said to be *explosive* if:

$$\forall \alpha \forall \beta(\Gamma, \alpha, \neg\alpha \Vdash \beta)$$

Thus, a theory is called explosive if it trivializes when exposed to a pair of contradictory formulas. Evidently, if a theory is trivial, then it is explosive by (Con2). Also, if a theory is contradictory and explosive, then it is trivial by (Con3).

The above definitions may be immediately upgraded from theories to logics. We will say that **L** is *contradictory* if all of its theories are contradictory, that is:

$$\forall\Gamma\exists\alpha(\Gamma \Vdash \alpha \text{ and } \Gamma \Vdash \neg\alpha)$$

In the same spirit, we will say that **L** is *trivial* if all of its theories are trivial, and **L** is *explosive* if all of its theories are explosive.

Because of the monotonicity property (Con2), it is clear that an S-logic **L** is contradictory / trivial / explosive if, and only if, its empty theory is contradictory / trivial / explosive.

We are now in position to give a formal definition for some *logical principles* as applied to a generic logic **L**:

Principle of Non-Contradiction (**L** is non-contradictory)

$$\exists\Gamma\forall\alpha(\Gamma \nVdash \alpha \text{ or } \Gamma \nVdash \neg\alpha) \tag{1}$$

Principle of Non-Triviality (**L** is non-trivial)

$$\exists\Gamma\exists\alpha(\Gamma \nVdash \alpha) \tag{2}$$

Principle of Explosion (**L** is explosive)

$$\forall\Gamma\forall\alpha\forall\beta(\Gamma, \alpha, \neg\alpha \Vdash \beta) \tag{3}$$

The last principle is also often referred to as *Pseudo-Scotus* or Principle of *Ex Contradictione Sequitur Quodlibet*.[4]

It is clear that the three principles are interrelated:

THEOREM 1.
(i) A trivial logic is both contradictory and explosive.
(ii) An explosive logic fails the Principle of Non-Triviality if, and only if, it fails the Principle of Non-Contradiction. ■

The logics disrespecting (1) are sometimes called *dialectical*. However, the immense majority of the paraconsistent logics in the literature (including the ones studied here) are *not* dialectical. Indeed, they usually have non-contradictory empty theories, and thus their axioms are non-contradictory, and their inference rules do not generate contradictions from these axioms. All paraconsistent logics which we will present here are in some sense more careful than classical logic, once they extract less consequences than classical logic extracts from the same given theory, or at most the

[4]In fact, single-conclusion logics as those we work with here cannot see the difference between *Pseudo-Scotus* and *Ex Contradictione*, but those principles can be sharply distinguished in a multiple-conclusion environment. Moreover, in such an environment, several forms of triviality, or *overcompleteness*, may be very naturally set apart (cf. [Marcos, 2005c] and [Marcos, 2007a]).

same set of consequences, but never more. The paraconsistent logics studied in the present chapter (as most paraconsistent logics in the literature) do not validate any bizarre form of reasoning, and do not beget contradictory consequences if such consequences were already not derived in classical logic.

2.2 Paraconsistency: Between inconsistency and triviality

As mentioned before, some decades ago, Stanisław Jaśkowski ([Jaśkowski, 1948]), David Nelson ([Nelson, 1959]), and Newton da Costa ([da Costa, 1963]), the founders of paraconsistent logic, proposed, independently, the study of logics which could accommodate contradictory yet non-trivial theories. For da Costa, a logic is *paraconsistent*[5] *with respect to* $\neg$ if it can serve as a basis for $\neg$-contradictory yet non-trivial theories, that is:

$$\exists\Gamma\exists\alpha\exists\beta(\Gamma \Vdash \alpha \text{ and } \Gamma \Vdash \neg\alpha \text{ and } \Gamma \nVdash \beta) \tag{4}$$

Notice that, in our present framework, the notion of a paraconsistent logic has, in principle, nothing to do with the rejection of the Principle of Non-Contradiction, as it is commonly held. On the other hand, it is intimately connected to the rejection of the Principle of Explosion. Indeed, Jaśkowski defined a $\neg$-paraconsistent logic as a logic in which (3) fails, that is:

$$\exists\Gamma\exists\alpha\exists\beta(\Gamma, \alpha, \neg\alpha \nVdash \beta) \tag{5}$$

Using (Con1) and (Con3) it is easy to prove that (4) and (5) are equivalent ways of defining a paraconsistent logic. Whenever it is clear from the context, we will omit the $\neg$ symbol and refer simply to *paraconsistent logics*.

It is very important to observe that a logic where all contradictions are equivalent cannot be paraconsistent. To understand that point it is convenient first to make precise the concept of equivalence between sets of formulas: Γ and Δ are said to be *equivalent* if

$$\forall\alpha \in \Delta(\Gamma \Vdash \alpha) \text{ and } \forall\alpha \in \Gamma(\Delta \Vdash \alpha)$$

In particular, we say that two formulas α and β are *equivalent* if the sets $\{\alpha\}$ and $\{\beta\}$ are equivalent, that is:

$$(\alpha \Vdash \beta) \text{ and } (\beta \Vdash \alpha)$$

We denote these facts by writing, respectively, $\Gamma \dashv\Vdash \Delta$, and $\alpha \dashv\Vdash \beta$. The equivalence between formulas is clearly an equivalence relation, because of (Con1) and (Con3). However, the equivalence between sets is not, in

[5] As a matter of fact, this appellation would be coined only in the 70s by the Peruvian philosopher Francisco Miró Quesada.

general, an equivalence relation, unless the following property holds in $\mathbf{L}$:

(Con6) $[\forall \beta \in \Delta(\Gamma \Vdash \beta)$ and $\Delta \Vdash \alpha]$ implies $\Gamma \Vdash \alpha$ (cut for sets)

Logics based on consequence relations that respect clauses (Con1), (Con2) and (Con6) will here be called (single-conclusion) *T-logics* (where T stands for *Tarskian*).

REMARK 2. (i) In logics defined by way of a collection of finite-valued truth-tables or by way of Hilbert calculi with schematic axioms and finitary rules, (Con1)–(Con6) all hold good. This is the case of most logics mentioned in the present paper.

(ii) (Con1) and (Con6) guarantee that $\dashv\Vdash$ defines an equivalence relation over sets of formulas.

(iii) Condition (Con3) follows from {(Con1), (Con2), (Con6)}. Indeed, suppose that (a) $\Delta \Vdash \alpha$ and (b) $\Gamma, \alpha \Vdash \beta$. By (Con1) we can further assume that (c) $\Delta, \Gamma \Vdash \gamma$, for every $\gamma \in \Gamma$. But if we apply (Con2) to hypothesis (a) it follows that (d) $\Delta, \Gamma \Vdash \alpha$. Using (Con6) on (c), (d) and (b) it follows that $\Delta, \Gamma \Vdash \beta$.

(iv) Condition (Con2) follows from {(Con1), (Con6)}. Indeed, suppose that (a) $\Delta \Vdash \alpha$ and (b) $\Delta \subseteq \Gamma$. From (b) and (Con1), we conclude that (c) $\Gamma \Vdash \delta$, for all $\delta \in \Delta$. Then, using (Con6) on (c) and (a) it follows that $\Gamma \Vdash \alpha$.

(v) Condition (Con3) follows from {(Con1), (Con6)}. To check that, compose (iii) and (iv).

(vi) Condition (Con6) does not follow from {(Con1), (Con2), (Con3)}. Indeed, consider for instance the logic $\mathbf{L}_{\mathbb{R}} = \langle \mathbb{R}, \Vdash \rangle$ such that $\mathbb{R}$ is the set of real numbers, and $\Vdash$ is defined as follows:

$$\begin{aligned} \Gamma \Vdash x \quad \text{iff} \quad & x \in \Gamma, \text{ or } x = \tfrac{1}{n} \text{ for some } n \in \mathbb{N},\ n \geq 1, \text{ or} \\ & \text{there is a sequence } (x_n)_{n\in\mathbb{N}} \text{ contained in } \Gamma \text{ such that} \\ & (x_n)_{n\in\mathbb{N}} \text{ converges to } x. \end{aligned}$$

It is easy to see that $\mathbf{L}_{\mathbb{R}}$ satisfies (Con1), (Con2) and (Con3). But (Con6) is not valid in $\mathbf{L}_{\mathbb{R}}$. Indeed, take $\Gamma = \varnothing$, $\Delta = \{\frac{1}{n} : n \in \mathbb{N},\ n \geq 1\}$ and $\alpha = 0$. Then the antecedent of (Con6) is true: Every element of Δ is a thesis, and Δ contains the sequence $(\frac{1}{n})_{n\in\mathbb{N}}$, which converges to 0. However, the consequent of (Con6) is false: 0 is not a thesis of $\mathbf{L}_{\mathbb{R}}$.

Observe, by the way, that in $\mathbf{L}_{\mathbb{R}}$ the relation $\dashv\Vdash$ between sets of formulas is not transitive: Take Δ as above, and consider $\Delta_0 = \{0\}$ and $\Delta_1 = \{1\}$. Then $\Delta_0 \dashv\Vdash \Delta$ and $\Delta \dashv\Vdash \Delta_1$, but it is not the case that $\Delta_0 \dashv\Vdash \Delta_1$, because $\Delta_1 \nVdash 0$.

Do remark that, as a particular consequence of the above items, T-logics may be seen as specializations of S-logics. ■

Most logics we will study in the present paper are natural examples of T-logics. For many proofs that will be presented below, however, the assumption of an S-logic will suffice.

THEOREM 3. Let $\mathbf{L}$ be a T-logic. Then, in case all contradictions are equivalent in $\mathbf{L}$, it follows that $\mathbf{L}$ is not paraconsistent.

Proof. Take an arbitrary set Γ in $\mathbf{L}$. Suppose that all contradictions are equivalent, that is, for arbitrary α and β, $\{\alpha, \neg\alpha\} \dashv\Vdash \{\beta, \neg\beta\}$. Then, using (Con2), $\Gamma \cup \{\alpha, \neg\alpha\}$ is β-contradictory for an arbitrary β, and in particular $\Gamma, \alpha, \neg\alpha \Vdash \beta$. ■

By contrapositive reasoning, the above theorem may be rephrased as stating the following: If a T-logic $\mathbf{L}$ is paraconsistent, then there exist pairs of non-equivalent contradictions in $\mathbf{L}$.

DEFINITION 4. The logic $\mathbf{L}$ is called *consistent* if it is both explosive and non-trivial, that is, if $\mathbf{L}$ respects both (3) and (2). $\mathbf{L}$ is called *inconsistent*, otherwise. ■

Paraconsistent logics are inconsistent, in that they control explosiveness, but they can do so in a variety of ways. Trivial logics are also inconsistent, by the above definition. What distinguishes a paraconsistent logic from a trivial logic is that a trivial logic does not disallow any inference: It accepts everything. As a consequence of the above definition of consistency, a third equivalent approach to the notion of paraconsistency may be proposed, parallel to those from definitions (4) and (5):

A logic is paraconsistent if it is inconsistent yet non-trivial. (6)

The compatibility of paraconsistency with the existence of some suitable explosive or trivial proper theories makes some paraconsistent logics able to recover classical reasoning, as we will see in Section 3.6. We will from now on introduce some specializations on the above definitions and principles.

A logic $\mathbf{L}$ is said to be *finitely trivializable* when it has finite trivial theories. Evidently, if a logic is explosive, then it is finitely trivializable. Non-explosive logics might be finitely trivializable or not.

A formula ξ in $\mathbf{L}$ is a *bottom particle* if it can, by itself, trivialize the logic, that is:

$$\forall\Gamma\forall\beta(\Gamma, \xi \Vdash \beta)$$

A bottom particle, when it exists, will here be denoted by $\bot$. This notation is unambiguous in the following sense: Any two bottom particles are equivalent. If in a given logic a bottom particle is also a thesis, then the logic is trivial — in which case, of course, all formulas turn out to be bottom particles.

The existence of bottom particles inside a given logic $\mathbf{L}$ is regulated by the following principle:

Principle of *Ex Falso Sequitur Quodlibet*

$$\exists\xi\forall\Gamma\forall\beta(\Gamma, \xi \Vdash \beta)(\mathbf{L} \text{ has a bottom particle}) \quad (7)$$

As it will be seen, the existence of logics that do not respect (3) while still respecting (7) (as all **LFI**s of the present chapter) shows that *ex contradictione* does not need to be identified with *ex falso*, contrary to what is commonly held in the literature.

The dual concept of a bottom particle is that of a *top particle*, that is, a formula ζ which follows from every theory:

$$\forall\Gamma(\Gamma \Vdash \zeta)$$

We will denote any fixed such particle, when it exists, by $\top$ (again, this notation is unambiguous). Evidently, given a logic, any of its theses will constitute such a top particle (and logics with no theses, like Kleene's 3-valued logic, have no such particles). It is easy to see that the addition of a top particle to a given theory is pretty innocuous, for in that case $\Gamma, \top \Vdash \alpha$ if and only if $\Gamma \Vdash \alpha$.

Henceforth, a formula φ of **L** constructed using all and only the variables $p_0, \ldots, p_n$ will be denoted by $\varphi(p_0, \ldots, p_n)$. This formula will be said to *depend only* on the variables that occur in it. The notation may be generalized to sets, and the result is denoted by $\Gamma(p_0, \ldots, p_n)$. If $\gamma_0, \ldots, \gamma_n$ are formulas then $\varphi(\gamma_0, \ldots, \gamma_n)$ will denote the (simultaneous) substitution of p_i by γ_i in $\varphi(p_0, \ldots, p_n)$ (for $i = 0, \ldots, n$). Given a set of formulas $\Gamma(p_0, \ldots, p_n)$, we will write $\Gamma(\gamma_0, \ldots, \gamma_n)$ with an analogous meaning.

DEFINITION 5. We say that a logic **L** has a *supplementing negation* if there is a formula $\varphi(p_0)$ such that:

(a) $\varphi(\alpha)$ is not a bottom particle, for some α;

(b) $\forall\Gamma\forall\alpha\forall\beta(\Gamma, \alpha, \varphi(\alpha) \Vdash \beta)$ ■

Observe that the same logic might have several non-equivalent supplementing negations (check Remark 43).

Consider a logic having a supplementing negation, and denote it by $\sim$. Parallel to the definition of contradictoriness with respect to $\neg$, we might now define a theory Γ to be *contradictory with respect to* $\sim$ if it is such that:

$$\exists\alpha(\Gamma \Vdash \alpha \text{ and } \Gamma \Vdash \sim\alpha)$$

Accordingly, a logic **L** could be said to be *contradictory with respect to* $\sim$ if all of its theories were contradictory with respect to $\sim$. Obviously, by design, no logic can be $\sim$-paraconsistent, or even $\sim$-contradictory, if $\sim$ is a supplementing negation, and a logic that has a supplementing negation must satisfy the Principle of Non-Contradiction with respect to this negation. The main logics studied in this paper are all endowed with supplementing negations. The availability of some specific supplementing negations makes some paraconsistent logics able to easily emulate classical negation (see Subsection 3.6).

Here we may of course introduce yet another variation on (3):

Supplementing Principle of Explosion

$$\mathbf{L} \text{ has a supplementing negation} \tag{8}$$

Supplementing negations are very common. We will show here some sufficient conditions for their definition. The presence of a convenient implication in our logics is often convenient so as to help explicitly internalizing the definition of new connectives.

DEFINITION 6. We say that a logic $\mathbf{L}$ has a *deductive implication* if there is a formula $\psi(p_0, p_1)$ such that:

(a) $\psi(\alpha, \beta)$ is not a bottom particle, for some choice of α and β;

(b) $\forall\alpha\forall\beta\forall\Gamma(\Gamma \Vdash \psi(\alpha, \beta)$ implies $\Gamma, \alpha \Vdash \beta)$;

(c) $\psi(\alpha, \beta)$ is not a top particle, for some choice of α and β;

(d) $\forall\alpha\forall\beta\forall\Gamma(\Gamma, \alpha \Vdash \beta$ implies $\Gamma \Vdash \psi(\alpha, \beta))$. ■

Inside the most usual logics, condition (b) is usually guaranteed by the validity of the rule of *modus ponens*, while condition (d) is guaranteed by the so-called 'deduction theorem' (when this theorem holds). Obviously, any logic having a deductive implication will be non-trivial, by condition (a).

THEOREM 7. Let $\mathbf{L}$ be a non-trivial logic endowed with a bottom particle $\bot$ and a deductive implication $\rightarrow$.
(i) Let $\neg$ be some negation symbol, and suppose that it satisfies:
(a) $\Gamma, \neg\alpha \Vdash \alpha \rightarrow \bot$;
(b) $\Gamma, \neg\alpha \rightarrow \bot \Vdash \alpha$.
Then, this $\neg$ is a supplementing negation.
(ii) Suppose, otherwise, that the following is the case:
(c) $\alpha \rightarrow \bot \nVdash \bot$, for some formula α.
Then, a supplementing negation may be defined by setting $\neg\alpha \stackrel{\text{def}}{=} \alpha \rightarrow \bot$.

Proof. Item (i). By hypothesis (a) and the properties of the bottom and the implication, we have $\Gamma, \alpha, \neg\alpha \Vdash \beta$. Now, suppose $\neg\alpha$ defines a bottom particle, for any choice of α. Then, by the deduction theorem, $\Gamma \Vdash \neg\alpha \rightarrow \bot$, for an arbitrary Γ. Thus, by (b) and (Con3), $\Gamma \Vdash \alpha$. But this cannot be the case, as $\mathbf{L}$ is non-trivial.
Item (ii) is a straightforward consequence of the above definitions, and we leave it as an exercise for the reader. ■

One might also consider the dual of a supplementing negation:

DEFINITION 8. We say that a logic $\mathbf{L}$ has a *complementing negation* if there is a formula $\psi(p_0)$ such that:

(a) $\psi(\alpha)$ is not a top particle, for some α;

(b) $\forall\Gamma\forall\alpha(\Gamma, \alpha \Vdash \psi(\alpha)$ implies $\Gamma \Vdash \psi(\alpha))$.

We say that $\mathbf{L}$ has a *classical negation* if it has some (primitive or defined) negation connective that is both supplementing and complementing. As a particular consequence of this definition, it can be easily checked that for any classical negation $\neg$ the equivalence $(\neg\neg\alpha \dashv\Vdash \alpha)$ will be derivable. ■

Yet some other versions of explosiveness can here be considered:

DEFINITION 9. Let $\mathbf{L}$ be a logic, and let $\sigma(p_0, \ldots, p_n)$ be a formula of $\mathbf{L}$. (i) We say that $\mathbf{L}$ is *partially explosive with respect to* σ, or *σ-partially explosive*, if:

(a) $\sigma(\beta_0, \ldots, \beta_n)$ is not a top particle, for some choice of $\beta_0, \ldots, \beta_n$;

(b) $\forall\Gamma\forall\beta_0 \ldots \forall\beta_n\forall\alpha(\Gamma, \alpha, \neg\alpha \Vdash \sigma(\beta_0, \ldots, \beta_n))$.

(ii) $\mathbf{L}$ is *boldly paraconsistent* if there is no σ such that $\mathbf{L}$ is σ-partially explosive.

(iii) $\mathbf{L}$ is said to be *controllably explosive in contact with* σ, if:

(a) $\sigma(\alpha_0, \ldots, \alpha_n)$ and $\neg\sigma(\alpha_0, \ldots, \alpha_n)$ are not bottom particles, for some choice of $\alpha_0, \ldots, \alpha_n$;

(b) $\forall\Gamma\forall\alpha_0 \ldots \forall\alpha_n\forall\beta(\Gamma, \sigma(\alpha_0, \ldots, \alpha_n), \neg\sigma(\alpha_0, \ldots, \alpha_n) \Vdash \beta)$. ■

EXAMPLE 10. A well-known example of a logic that is not explosive but is partially explosive, is provided by Kolmogorov & Johánsson's Minimal Intuitionistic Logic, *MIL*, obtained by the addition to the positive fragment of intuitionistic logic (see Remark 29 below) of some weak forms of *reductio ad absurdum* (cf. [Johánsson, 1936] and [Kolmogorov, 1967]). In this logic, the intuitionistically valid inference $(\Gamma, \alpha, \neg\alpha \Vdash \beta)$ fails, but $(\Gamma, \alpha, \neg\alpha \Vdash \neg\beta)$ holds good. This means that *MIL* is paraconsistent, but not boldly paraconsistent, as all negated propositions can be inferred from any given contradiction. A class of (obviously non-boldly) paraconsistent logics extending *MIL* is studied in [Odintsov, 2005]. ■

The requirement that a paraconsistent logic should be boldly paraconsistent was championed by [Urbas, 1990]. The class of boldly paraconsistent logics is surely very natural and pervasive. From now on, we will be making an effort, as a matter of fact, to square our paraconsistent logics into this class (check Theorems 20, 38 and 130).

Most paraconsistent logics studied in this chapter are also controllably explosive (check, in particular, Theorem 79, but a particularly strong counterexample may be found in Example 17).

We should observe that conjunction may play a central role in relating contradictoriness and triviality.

DEFINITION 11. A logic **L** is said to be *left-adjunctive* if there is a formula $\psi(p_0, p_1)$ such that:

(a) $\psi(\alpha, \beta)$ is not a bottom particle, for some α and β;

(b) $\forall\alpha\forall\beta\forall\Gamma\forall\gamma(\Gamma, \alpha, \beta \Vdash \gamma$ implies $\Gamma, \psi(\alpha, \beta) \Vdash \gamma)$. ■

The formula $\psi(\alpha, \beta)$, when it exists, will often be denoted by $(\alpha \wedge \beta)$, and the sign $\wedge$ will be called a *left-adjunctive conjunction* (but it will not necessarily have, of course, all properties of a classical conjunction). Similarly, we can define the following:

DEFINITION 12. A logic **L** is said to be *left-disadjunctive* if there is a formula $\varphi(p_0, p_1)$ such that:

(a) $\varphi(\alpha, \beta)$ is not a top particle, for some α and β;

(b) $\forall\alpha\forall\beta\forall\Gamma\forall\gamma(\Gamma, \varphi(\alpha, \beta) \Vdash \gamma$ implies $\Gamma, \alpha, \beta \Vdash \gamma)$. ■

In general, whenever there is no risk of misunderstanding or of misidentification of different entities, we might also denote the formula $\varphi(\alpha, \beta)$, when it exists, by $(\alpha \wedge \beta)$, and we will accordingly call $\wedge$ a *left-disadjunctive conjunction*. Of course, a logic can have just one of these conjunctions, or it can have both a left-adjunctive conjunction and a left-disadjunctive conjunction without the two of them coinciding. In natural deduction, clause (b) of Definition 11 corresponds to conjunction elimination, and clause (b) of Definition 12 corresponds to conjunction introduction.

It is straightforward to prove the following:

THEOREM 13. Let **L** be a left-adjunctive logic. (i) If **L** is finitely trivializable (in particular, if it has a supplementing negation), then it has a bottom particle. (ii) If **L** respects *ex contradictione*, then it also respects *ex falso*. ■

EXAMPLE 14. The 'pre-discussive' logic J proposed in [Jaśkowski, 1948], in the usual signature of classical logic, is such that:

$$\Gamma \Vdash_J \alpha \text{ iff } \Diamond\Gamma \Vdash_{S5} \Diamond\alpha,$$

where $\Diamond\Gamma = \{\Diamond\gamma : \gamma \in \Gamma\}$, $\Diamond$ denotes the possibility operator, and $\Vdash_{S5}$ denotes the consequence relation defined by the well-known modal logic $S5$. It is easy to see that $(\alpha, \neg\alpha \Vdash_J \beta)$ does not hold in general, though $(\alpha \wedge \neg\alpha) \Vdash_J \beta$ does hold good, for any formulas α and β. This phenomenon can only happen because J is left-adjunctive but not left-disadjunctive. Hence, Theorem 13 still holds for J, but this logic provides a simple example of a logic that respects the Principle of *Ex Falso Sequitur Quodlibet* (7) but not the Principle of *Ex Contradictione Sequitur Quodlibet* (3). ■

The literature on paraconsistency (cf. section 4.2 of [Priest, 2002]) traditionally calls *non-adjunctive* the logics failing left-disadjunctiveness. In

the present paper, conjunctions that are both left-adjunctive and left-dis-adjunctive will be called *standard*.

3 LFIS AND THEIR RELATIONSHIP TO CLASSICAL LOGIC

3.1 Introducing LFIs and C-systems

From now on, we will concentrate on logics which are paraconsistent but nevertheless have some special explosive theories, as those discussed in the last section. With the help of such theories some concepts can be studied under a new light — this is the case of the notion of *consistency* (and its opposite, the notion of *inconsistency*), as we shall see. This section will introduce the Logics of Formal Inconsistency as the paraconsistent logics that respect a certain Gentle Principle of Explosion, to be clarified below. By way of motivation, we start with a few helpful definitions and concrete examples.

Given two logics $\mathbf{L}1 = \langle For_1, \Vdash_1 \rangle$ and $\mathbf{L}2 = \langle For_2, \Vdash_2 \rangle$, we will say that $\mathbf{L}2$ is a (proper) *linguistic extension* of $\mathbf{L}1$ if For_1 is a (proper) subset of For_2, and we will say that $\mathbf{L}2$ is a (proper) *deductive extension* of $\mathbf{L}1$ if $\Vdash_1$ is a (proper) subset of $\Vdash_2$. Finally, if $\mathbf{L}2$ is both a linguistic extension and a deductive extension of $\mathbf{L}1$, and if the restriction of $\mathbf{L}2$'s consequence relation $\Vdash_2$ to the set For_1 will make it identical to $\Vdash_1$ (that is, if $For_1 \subseteq For_2$, and for any $\Gamma \cup \{\alpha\} \subseteq For_1$ we have that $\Gamma \Vdash_2 \alpha$ iff $\Gamma \Vdash_1 \alpha$) then we will say that $\mathbf{L}2$ is a *conservative extension* of $\mathbf{L}1$ (and similarly for *proper* conservative extensions). In any of the above cases we can more generally say that $\mathbf{L}2$ is an *extension* of $\mathbf{L}1$, or that $\mathbf{L}1$ is a *fragment* of $\mathbf{L}2$. These concepts will be used here to compare a number of logics that will be presented. Most paraconsistent logics in the literature, and all of those studied here, are proper deductive fragments of classical logic written in a convenient signature.

REMARK 15. From here on, Σ will denote the signature containing the binary connectives $\wedge$, $\vee$, $\rightarrow$, and the unary connective $\neg$, such that $\mathcal{P} = \{p_n : n \in \omega\}$ is the set of atomic formulas. By *For* we will denote the set of formulas freely generated by $\mathcal{P}$ over Σ.

In the same spirit, Σ° will denote the signature obtained by the addition to Σ of a new unary connective $\circ$ to the signature Σ, and For° will denote the algebra of formulas for the signature Σ°. ■

DEFINITION 16. A *many-valued semantics* for a set of formulas *For* will here be any collection Sem of mappings v_k: $For \longrightarrow \mathcal{V}_k$, called *valuations*, where the set of *truth-values* in $\mathcal{V}_k$ is separated into *designated* values $\mathcal{D}_k$ (denoting the set of 'true values') and *undesignated* values $\mathcal{U}_k$ (denoting the set of 'false values'), that is, $\mathcal{V}_k$ is such that $\mathcal{V}_k = \mathcal{D}_k \cup \mathcal{U}_k$ and $\mathcal{D}_k \cap \mathcal{U}_k = \varnothing$, for each $v \in \mathsf{Sem}$. A (truth-preserving single-conclusion) many-valued

entailment relation $\models_{\mathsf{Sem}} \subseteq \wp(For) \times For$ can then be defined by setting, for every choice of $\Gamma \cup \{\alpha\} \subseteq For$:

$$\Gamma \models_{\mathsf{Sem}} \alpha \text{ iff, for every } v \in \mathsf{Sem},\ v(\alpha) \in \mathcal{D} \text{ whenever } v(\Gamma) \subseteq \mathcal{D}.$$

A nice general abstract result can be proven to the effect that a consequence relation characterizes a T-logic (recall Subsection 2.2) if, and only if, it is determined by a many-valued entailment relation (check [Marcos, 2004; Caleiro *et al.*, 2005a], and the references therein). A distinguished class of many-valued semantics that will be much explored in the present paper, starting from Subsection 3.3, is the class of semantics in which $\mathcal{D}$ and $\mathcal{U}$ are fixed singletons (representing 'truth' and 'falsity') throughout every $v \in \mathsf{Sem}$. Those semantics are now known as *bivaluation semantics*.

A very usual particular class of many-valued semantics is the class of *truth-functional semantics*, which include those many-valued semantics such that $\mathcal{V}$, $\mathcal{D}$ and $\mathcal{U}$ are fixed sets of truth-values throughout every $v \in \mathsf{Sem}$, and such that the truth-values are organized into an algebra similar to the algebra of formulas, that is, for every κ-ary connective in the signature Σ that defines For there is a corresponding κ-ary operator over $\mathcal{V}$, where κ is the cardinality of $\mathcal{V}$. In case $\kappa < \omega$ we say that we are talking about a *finite-valued truth-functional logic*.

We will often present truth-functional T-logics below simply in terms of sets of truth-tables and corresponding designated values defining the behavior of the connectives from the signature, and take it for granted that the reader assumes that and understands how those tables characterize an entailment relation $\models$, defined as above. Not all logics, and not all paraconsistent logics, have truth-functional semantics, though. Partially explosive paraconsistent logics such as *MIL* (check Example 10) provide indeed prime examples of logics that are not characterizable by truth-functional semantics, neither finite-valued nor infinite-valued (for a discussion on that phenomenon, check [Marcos, 2007b], and the references therein).

Some useful generalizations of truth-functional semantics include *nondeterministic semantics* and *possible-translations semantics* based on truth-functional many-valued logics (presented below, starting from Subsection 3.4). ■

EXAMPLE 17. Consider the logic presented by way of the following truth-tables:

$\wedge$	1	$^1/_2$	0
1	1	$^1/_2$	0
$^1/_2$	$^1/_2$	$^1/_2$	0
0	0	0	0

$\vee$	1	$^1/_2$	0
1	1	1	1
$^1/_2$	1	$^1/_2$	$^1/_2$
0	1	$^1/_2$	0

$\rightarrow$	1	$^1/_2$	0
1	1	$^1/_2$	0
$^1/_2$	1	$^1/_2$	0
0	1	1	1

	$\neg$
1	0
$^1/_2$	$^1/_2$
0	1

where both 1 and $\frac{1}{2}$ are designated values. *Pac* is the name under which this logic appeared in [Avron, 1991] (Section 3.2.2), though it had previously

appeared, for instance, in [Avron, 1986], under the denomination $RM_3^{\widetilde{\supset}}$, and, even before that, in [Batens, 1980], where it was called PI^s. The logic Pac conservatively extends the logic LP by the addition of a classical implication. LP is an early example of a 3-valued paraconsistent logic with classic-like operators for a standard conjunction and a standard disjunction, and it was introduced in [Asenjo, 1966] and investigated in [Priest, 1979].

In Pac, for no formula α it is the case that $\alpha, \neg\alpha \vdash_{Pac} \beta$ for all β. So, Pac is not a controllably explosive logic. A classical negation for Pac would be illustrated by the truth-table:

	$\sim$
1	0
$^1/_2$	0
0	1

However, it should be clear that such a negation is *not* definable in Pac. Indeed, any truth-function of this logic having only $\frac{1}{2}$'s as input will also have $\frac{1}{2}$ as output. As a consequence, Pac has no bottom particle (and this logic also cannot express the consistency of its formulas, as we shall see below). Being a left-adjunctive logic as well, Pac is, consequently, not finitely trivializable. ■

EXAMPLE 18. In adding to Pac either a supplementing negation as above or a bottom particle, one obtains a well-known conservative extension of it, obviously still paraconsistent, but this time a logic that has some interesting explosive theories: It satisfies, in particular, principles (7) and (8) from the previous subsection. This logic was introduced in [Schütte, 1960] for proof-theoretical reasons and independently investigated under the appellation $\mathbf{J}_3$ in [D'Ottaviano and da Costa, 1970] as a 'possible solution to the problem of Jaśkowski'. It also reappeared quite often in the literature after that, for instance as the logic **CLuNs** in [Batens and De Clercq, 2000]. In [D'Ottaviano and da Costa, 1970]'s first presentation of $\mathbf{J}_3$, a 'possibility connective' ∇ was introduced instead of the supplementing negation $\sim$. In [Epstein, 2000] this logic was reintroduced having also a sort of 'consistency connective' $\circ$ (originally denoted by ©) as primitive. The truth-tables of ∇ and $\circ$ are as follows:

	∇	$\circ$
1	1	1
$^1/_2$	1	0
0	0	1

The expressive and inferential power of this logic was more deeply explored in [Avron, 1999] and in [Carnielli *et al.*, 2000]. The latter paper also explores

the possibility of applying this logic to the study of inconsistent databases (for a more technical perspective check [de Amo *et al.*, 2002]), abandoning $\sim$ and ∇ but still retaining $\circ$ as primitive. This logic (renamed **LFI1** in the signature Σ°) has been argued to be appropriate for formalizing the notion of consistency in a very convenient way, as discussed below. It is worth noticing that $\sim\alpha$ and $\nabla\alpha$ may be defined in **LFI1** as $(\neg\alpha \wedge \circ\alpha)$ and $(\alpha \vee \neg\circ\alpha)$, respectively. Alternatively, $\circ\alpha \stackrel{\text{def}}{=} (\neg\nabla\alpha \vee \neg\nabla\neg\alpha)$. A complete axiomatization for **LFI1** is presented in Theorem 127. ■

EXAMPLE 19. Paraconsistency and many-valuedness have often been companions. In [Sette, 1973] the following 3-valued logic, alias $\mathbf{P}^1$, was studied:

$\wedge$	1	$^1/_2$	0
1	1	1	0
$^1/_2$	1	1	0
0	0	0	0

$\vee$	1	$^1/_2$	0
1	1	1	1
$^1/_2$	1	1	1
0	1	1	0

$\rightarrow$	1	$^1/_2$	0
1	1	1	0
$^1/_2$	1	1	0
0	1	1	1

	$\neg$
1	0
$^1/_2$	1
0	1

where 1 and $\frac{1}{2}$ are the designated values. The truth-table of the consistency connective $\circ$ as in Example 18 can now be defined via $\circ\alpha \stackrel{\text{def}}{=} \neg\neg\alpha \vee \neg(\alpha \wedge \alpha)$. The logic $\mathbf{P}^1$ has the remarkable property of being controllably explosive in contact with arbitrary non-atomic formulas, that is, the paraconsistent behavior obtains only at the atomic level: $\alpha, \neg\alpha \vDash \beta$, for arbitrary non-atomic α. Moreover, another property of this logic is that $\vDash \circ\alpha$ holds for non-atomic α. Those two properties are in fact not related by a mere accident, but as an instance of Theorem 79. A complete axiomatization for the logic $\mathbf{P}^1$ is presented in Theorem 127. ■

We had committed ourselves to present paraconsistent logics that would be boldly paraconsistent (recall Definition 9(ii)). The logics from Examples 18 and 19 can indeed be seen to enjoy this property:

THEOREM 20. **LFI1** and $\mathbf{P}^1$ are boldly paraconsistent. And so are their fragments.

Proof. Assume $\Gamma \nvdash \sigma(p_0, \ldots, p_n)$ for some appropriate choice of formulas. In particular, by (Con2), it follows that $\nvdash \sigma(p_0, \ldots, p_n)$. Now, consider a variable p not in $p_0, \ldots, p_n$. Let p be assigned the value $\frac{1}{2}$, and extend this assignment to the variables $p_0, \ldots, p_n$ so as to give the value 0 to $\sigma(p_0, \ldots, p_n)$. It is obvious that, in this situation, $p, \neg p \nvdash \sigma(p_0, \ldots, p_n)$. ■

Paraconsistent logics are tools for reasoning under conditions which do not presuppose consistency. If we understand consistency as what might be lacking to a contradiction for it to become explosive, logics like **LFI1** and $\mathbf{P}^1$ are clearly able to express the consistency (of a formula) at the object-language level. This feature will permit consistent reasoning to be recovered from inside an inconsistent environment.

In formal terms, consider a (possibly empty) set $\bigcirc(p)$ of formulas which depends only on the propositional variable p, satisfying the following: There are formulas α and β such that

$$\begin{array}{ll} \text{(a)} & \bigcirc(\alpha), \alpha \nvdash \beta; \\ \text{(b)} & \bigcirc(\alpha), \neg\alpha \nvdash \beta. \end{array}$$

We will call a theory Γ *gently explosive* (*with respect to* $\bigcirc(p)$) if:

$$\forall\alpha\forall\beta(\Gamma, \bigcirc(\alpha), \alpha, \neg\alpha \Vdash \beta).$$

A theory Γ will be said to be *finitely gently explosive* when it is gently explosive with respect to a finite set $\bigcirc(p)$.

A logic **L** will be said to be *(finitely) gently explosive* when there is a (finite) set $\bigcirc(p)$ such that all of the theories of **L** are (finitely) gently explosive (with respect to $\bigcirc(p)$). Notice that a finitely gently explosive theory is finitely trivialized in a very distinctive way.

We may now consider the following 'gentle' variations on the Principle of Explosion:

Gentle Principle of Explosion

$$\mathbf{L} \text{ is gently explosive with respect to some set } \bigcirc(p) \tag{9}$$

Finite Gentle Principle of Explosion

$$\mathbf{L} \text{ is gently explosive with respect to some finite set } \bigcirc(p) \tag{10}$$

For any formula α, the set $\bigcirc(\alpha)$ is intended to express, in a specific sense, the consistency of α relative to the logic **L**. When this set is a singleton, we will denote by $\circ\alpha$ the sole element of $\bigcirc(\alpha)$, and in this case $\circ$ defines a *consistency connective* or *consistency operator*. It is worth noting, however, that $\circ$ is not necessarily a primitive connective of the signature of **L**. In fact, several logics that will be studied below (namely, the so-called 'direct **dC**-systems', see Definition 32) present $\circ$ as a connective that is defined in terms of other connectives of a less complex underlying signature.

The above definitions are very natural, and paraconsistent logics with a consistency connective are in fact quite common. One way of seeing that is through the use of a classic-like (in fact, intuitionistic-like) disjunction:

DEFINITION 21. We say that a logic **L** *has a standard disjunction* if there is a formula $\psi(p_0, p_1)$ such that:

(a) $\psi(\alpha, \beta)$ is not a bottom particle, for some α and β;

(b) $\forall\alpha\forall\beta\forall\Gamma\forall\Delta\forall\gamma(\Gamma, \alpha \Vdash \gamma$ and $\Delta, \beta \Vdash \gamma$ implies $\Gamma, \Delta, \psi(\alpha, \beta) \Vdash \gamma)$;

(c) $\psi(\alpha, \beta)$ is not a top particle, for some α and β;

(d) $\forall\alpha\forall\beta\forall\Gamma\forall\gamma(\Gamma, \psi(\alpha, \beta) \Vdash \gamma$ implies $\Gamma, \alpha \Vdash \gamma$ and $\Gamma, \beta \Vdash \gamma)$.

In natural deduction, clause (b) corresponds to disjunction elimination, and clause (d) to disjunction introduction. The reader can now easily check that:

THEOREM 22. (i) Any non-trivial explosive theory / logic is finitely gently explosive, supposing that there is some formula α such that $\neg\alpha$ is not a bottom particle. (ii) Any left-adjunctive finitely gently explosive logic respects *ex falso*. (iii) Let $\mathbf{L}$ be a logic containing a bottom particle $\bot$, a standard disjunction $\vee$, an implication $\rightarrow$ respecting *modus ponens* and a negation $\neg$ such that there exists some formula α satisfying:
(a) $\alpha, (\neg\alpha \rightarrow \bot) \nvdash \bot$;
(b) $\neg\alpha, (\alpha \rightarrow \bot) \nvdash \bot$.
Then $\mathbf{L}$ defines a consistency operator $\circ\alpha \stackrel{\text{def}}{=} (\alpha \rightarrow \bot) \vee (\neg\alpha \rightarrow \bot)$. ■

We now define the Logics of Formal Inconsistency as the paraconsistent logics that can 'talk about consistency' in a meaningful way.

DEFINITION 23. A *Logic of Formal Inconsistency* (**LFI**) is any gently explosive paraconsistent logic, that is, any logic in which explosion, (3), does not hold, while gentle explosion, (9), holds good. ■

In other words, a logic $\mathbf{L}$ is an **LFI** (with respect to a negation $\neg$) if:

(a) $\exists\Gamma\exists\alpha\exists\beta(\Gamma, \alpha, \neg\alpha \nvdash \beta)$, and

(b) there exists a set of formulas $\bigcirc(p)$ depending exactly on the propositional variable p such that $\forall\Gamma\forall\alpha\forall\beta(\Gamma, \bigcirc(\alpha), \alpha, \neg\alpha \vdash \beta)$.

Besides the 3-valued paraconsistent logics presented in the above examples, we will study in this chapter several other paraconsistent logics based on different kinds of semantics. Many will have been originally proposed without a primitive consistency connective, but, being sufficiently expressive, they will often be shown to admit of such a connective. Examples of that phenomenon were already presented above, for the cases of **LFI1** and $\mathbf{P}^1$. Another interesting and maybe even surprising example of that phenomenon is provided by Jaśkowski's Discussive Logic **D2** (cf. [Jaśkowski, 1948] and [Jaśkowski, 1949]), the first paraconsistent logic ever to be introduced as such in the literature:

EXAMPLE 24. Let $\Sigma^{\Diamond}$ be the extension of the signature Σ obtained by the addition of a new unary connective $\Diamond$, and let $For^{\Diamond}$ be the corresponding algebra of formulas. Let $\vdash_{S5}$ be the consequence relation of modal logic *S5* over the language $For^{\Diamond}$. Consider a mapping $*$: $For \longrightarrow For^{\Diamond}$ such that:

1. $p^* = p$ for every $p \in \mathcal{P}$;
2. $(\neg\alpha)^* = \neg\alpha^*$;
3. $(\alpha \vee \beta)^* = \alpha^* \vee \beta^*$;
4. $(\alpha \wedge \beta)^* = \alpha^* \wedge \Diamond\beta^*$;

5. $(\alpha \rightarrow \beta)^* = \Diamond\alpha^* \rightarrow \beta^*$.

Given $\Gamma \subseteq For$, let Γ^* denote the subset $\{\alpha^* : \alpha \in \Gamma\}$ of $For^\Diamond$. For any $\Gamma \subseteq For^\Diamond$ let $\Diamond\Gamma = \{\Diamond\alpha : \alpha \in \Gamma\}$. Jaśkowski's Discussive logic **D2** is defined over the signature Σ as follows: $\Gamma \Vdash_{\mathbf{D2}} \alpha$ iff $\Diamond(\Gamma^*) \Vdash_{S5} \Diamond(\alpha^*)$, for any $\Gamma \cup \{\alpha\} \subseteq For$. Equivalently, **D2** may be introduced with the help of the pre-discussive logic J (recall Example 14), by setting $\Gamma \Vdash_{\mathbf{D2}} \alpha$ iff $\Gamma^* \Vdash_J \alpha^*$. With such definitions, **D2** can easily be seen to be non-explosive with respect to the negation $\neg$, that is, **D2** is paraconsistent (with respect to $\neg$). Consider now the following abbreviations defined on the set For (here, $\alpha \in For$):

$\top \stackrel{\text{def}}{=} (\alpha \vee \neg\alpha)$;
$\bot \stackrel{\text{def}}{=} \neg\top$;
$\blacksquare\alpha \stackrel{\text{def}}{=} (\neg\alpha \rightarrow \bot)$;
$\blacklozenge\alpha \stackrel{\text{def}}{=} \neg\blacksquare\neg\alpha$;
$\circ\alpha \stackrel{\text{def}}{=} (\blacklozenge\alpha \rightarrow \blacksquare\alpha)$.

It is an easy task to check now (say, using a Kripke semantics or tableaux for the logic $S5$) that in **D2** the formulas $\top$ and $\bot$ denote top and bottom particles, respectively, and $\circ$ behaves as a consistency operator (giving rise to gentle explosion). ■

THEOREM 25.
(i) Classical logic is not an **LFI**.

(ii) *Pac* (see Example 17) is also not an **LFI**.

(iii) **LFI1** (see Example 18) is an **LFI**.

(iv) $\mathbf{P}^1$ (see Example 19) is an **LFI**.

(v) Jaśkowski's Discussive Logic **D2** (see Example 24) is an **LFI**.

Proof. For item (i), note that explosion, (3), holds classically.

To check item (ii), let p be an atomic formula and let $\bigcirc(p)$ be the set of all formulas of *Pac* that depend only on p. The valuation from the truth-table that assigns $\frac{1}{2}$ to p and 0 to q is a model for $\bigcirc(p), p, \neg p$ but it invalidates gentle explosion (on q).

For item (iii), take consistency to be expressed in $\mathbf{J}_3$ by the connective $\circ$, as intended, that is, take $\bigcirc(\alpha) = \{\circ\alpha\}$. Obviously, $\bigcirc(\alpha), \alpha, \neg\alpha \vDash \beta$ holds. Take now a valuation from the truth-table that assigns 1 to p and notice that $\bigcirc(p), p \nvDash \beta$. Finally, take a valuation that assigns 0 to p and notice that $\bigcirc(p), \neg p \nvDash \beta$.

To check item (iv), again take consistency to be expressed in $\mathbf{P}^1$ by $\circ$ and note that $p, \neg p \nvDash q$, for atomic and distinct p and q.

Item (v) may be verified directly from the definitions in Example 24. ■

In accordance with definition (6) from Subsection 2.2, paraconsistent logics are the non-trivial logics whose negation fails the 'consistency presupposition'. Some inferences that depend on this presupposition, thus, will necessarily be lost. However, one might well expect that, if a sufficient number of 'consistency assumptions' are made, then those same inferences should be recovered. In fact, the **LFI**s are intended to be exactly the logics that can internalize this idea. To be more precise, and following [Marcos, 2005e]:

REMARK 26. Consider a logic $\mathbf{L}1 = \langle For_1, \Vdash_1 \rangle$ in which explosion holds good for a negation $\neg$, that is, a logic that satisfies, in particular, the rule $(\alpha, \neg\alpha \Vdash_1 \beta)$. Let $\mathbf{L}2 = \langle For_2, \Vdash_2 \rangle$ now be some other logic written in the same signature as $\mathbf{L}1$ such that: (i) $\mathbf{L}2$ is a proper deductive fragment of $\mathbf{L}1$ that validates inferences of $\mathbf{L}1$ only if they are compatible with the *failure* of explosion; (ii) $\mathbf{L}2$ is *expressive* enough so as to be an **LFI**, therefore, in particular, there will be in $\mathbf{L}2$ a set of formulas $\bigcirc(p)$ such that $(\bigcirc(\alpha), \alpha, \neg\alpha \Vdash_2 \beta)$ holds good; (iii) $\mathbf{L}1$ can in fact be *recovered* from $\mathbf{L}2$ by the addition of $\bigcirc(\alpha)$ as a new set of valid schemas / axioms. These constraints alone suggest that the reasoning of $\mathbf{L}1$ might somehow be recovered from inside $\mathbf{L}2$, if only a sufficient number of 'consistency assumptions' are added in each case. Thus, typically the following *Derivability Adjustment Theorem* (**DAT**) may be proven (as in [Marcos, 2005e]):

$$\forall\Gamma\forall\gamma\exists\Delta(\Gamma \Vdash_1 \gamma \text{ iff } \bigcirc(\Delta), \Gamma \Vdash_2 \gamma).$$

The **DAT** shows how the weaker logic $\mathbf{L}2$ can be used to 'talk about' the stronger logic $\mathbf{L}1$. The essential intuition behind such theorem was emphasized in [Batens, 1989], but an early version of that very idea can already be found in [da Costa, 1963] and [da Costa, 1974] (check our Theorem 112). On those grounds, **LFI**s are thus proposed and understood as the non-trivial inconsistent logics that can recover consistent inferences through convenient derivability adjustments. We will come back to this idea in Subsection 3.6 and Theorems 96, 112 and 113. ■

To get a bit more concrete, and at the same time specialize from the broad Definition 23 of **LFI**s, we introduce now the concept of a **C**-system.

DEFINITION 27. Let $\mathbf{L}1$ and $\mathbf{L}2$ be two logics defined over signatures Σ_1 and Σ_2, respectively, such that Σ_2 extends Σ_1, and Σ_2 contains a unary negation connective $\neg$ that does not belong to Σ_1. We say that $\mathbf{L}2$ is a **C**-*system based on* $\mathbf{L}1$ *with respect to* $\neg$ (in short, a **C**-*system*) if:

(a) $\mathbf{L}2$ is a conservative extension of $\mathbf{L}1$,
(b) $\mathbf{L}2$ is an **LFI** (with respect to $\neg$), such that the set $\bigcirc(p)$ is a singleton $\{\circ p\}$, that is, consistency may be defined as a formula $\varphi(p)$ in $\mathbf{L}2$,[6]

[6] In particular, $\varphi(p)$ could be of the form $\divideontimes(p)$ for $\divideontimes$ a unary connective of Σ_2.

(c) the non-explosive negation $\neg$ cannot be defined in **L**1,
(d) **L**1 is non-trivial. ■

All **C**-systems we will be studying below are examples of non-contradictory $\neg$-paraconsistent logical systems. Furthermore, they are equipped with supplementing negations and bottom particles, and they are based on classical propositional logic (in a convenient signature which includes an explicit connective for classical negation). Accordingly, they will all respect Principles (1), (2), (7), (8) and (9), but they will obviously disrespect (3).

As it will be seen in the following, the hierarchy of logics C_n, $1 \leq n < \omega$ (cf. [da Costa, 1963] or [da Costa, 1974]) provide clear illustrations of **C**-systems based on classical logic, provided that each C_n is presented in an extended signature including a connective for classical negation. The cautious reader should bear in mind that C_ω (cf. Definition 40 below), the logic proposed as a kind of 'limit' for the hierarchy is *not* a **C**-system, not even an **LFI**. The real deductive limit for the hierarchy, the logic C_{Lim}, is an interesting example of a gently explosive **LFI** that is not finitely so, and it was studied in [Carnielli and Marcos, 1999]. The next definition will recall the hierarchy C_n, $1 \leq n < \omega$, in an axiomatic formulation of our own:

DEFINITION 28. Recall, once more, the signature Σ from Remark 15. For every formula α, let $\circ\alpha$ be an abbreviation for the formula $\neg(\alpha \wedge \neg\alpha)$. The logic $C_1 = \langle For, \vdash_{C_1} \rangle$ may be axiomatized by the following schemas of a Hilbert calculus:

Axiom schemas:

(Ax1) $\alpha \rightarrow (\beta \rightarrow \alpha)$

(Ax2) $(\alpha \rightarrow \beta) \rightarrow ((\alpha \rightarrow (\beta \rightarrow \gamma)) \rightarrow (\alpha \rightarrow \gamma))$

(Ax3) $\alpha \rightarrow (\beta \rightarrow (\alpha \wedge \beta))$

(Ax4) $(\alpha \wedge \beta) \rightarrow \alpha$

(Ax5) $(\alpha \wedge \beta) \rightarrow \beta$

(Ax6) $\alpha \rightarrow (\alpha \vee \beta)$

(Ax7) $\beta \rightarrow (\alpha \vee \beta)$

(Ax8) $(\alpha \rightarrow \gamma) \rightarrow ((\beta \rightarrow \gamma) \rightarrow ((\alpha \vee \beta) \rightarrow \gamma))$

(Ax9) $\alpha \vee (\alpha \rightarrow \beta)$

(Ax10) $\alpha \vee \neg\alpha$

(Ax11) $\neg\neg\alpha \rightarrow \alpha$

(bc1) $\circ\alpha \rightarrow (\alpha \rightarrow (\neg\alpha \rightarrow \beta))$

(ca1) $(\circ\alpha \wedge \circ\beta) \rightarrow \circ(\alpha \wedge \beta)$

(ca2) $(\circ\alpha \wedge \circ\beta) \rightarrow \circ(\alpha \vee \beta)$

(ca3) $(\circ\alpha \wedge \circ\beta) \rightarrow \circ(\alpha \rightarrow \beta)$

Inference rule:

(MP) $\dfrac{\alpha,\ \alpha \rightarrow \beta}{\beta}$

In general, given a set of axioms and rules of a logic $\mathbf{L}$, we write $\Gamma \vdash_{\mathbf{L}} \alpha$ to say that there is proof in $\mathbf{L}$ of α from the premises in Γ. The subscript will be omitted when obvious from the context. If Γ is empty we say that α is a *theorem* of $\mathbf{L}$.

The logic C_1 is a **LFI** such that $\bigcirc(p) = \{\circ p\} = \{\neg(p \wedge \neg p)\}$. We shall see that axioms (bc1), and (ca1)–(ca3) can be stated in a new fashion by taking $\circ$ as a primitive connective instead of as an abbreviation. From these new axioms different logics will emerge. Moreover, since it is possible to define a classical negation $\sim$ in C_1 (namely, $\sim\alpha = \neg\alpha \wedge \circ\alpha$), this logic may be rewritten in an extended signature which contains $\sim$ as a primitive connective (and adding the obvious axioms identifying $\sim\alpha$ with $\neg\alpha \wedge \circ\alpha$), and so it is easy to see that C_1 (presented in the extended signature) is a **C**-system based on classical logic (see Remark 29 below).

Let α^1 abbreviate the formula $\neg(\alpha \wedge \neg\alpha)$, and α^{n+1} abbreviate the formula $(\neg(\alpha^n \wedge \neg\alpha^n))^1$. Then, each logic C_n of the hierarchy $\{C_n\}_{1 \leq n < \omega}$ may be obtained by assuming $\bigcirc(p) = \{p^1, \ldots, p^n\}$. This is equivalent, of course, to setting $\circ\alpha \stackrel{\text{def}}{=} \alpha^1 \wedge \ldots \wedge \alpha^n$ in axioms (bc1) and (ca1)–(ca3). It is immediate to see that every logic C_n is an **LFI**. Moreover, by considering the definable classical negation $\sim$ as a primitive connective, each C_n (presented in the extended signature) is a **C**-system based on classical logic. It is well known that each C_n properly extends each C_{n+1}. ■

REMARK 29. Let the signature Σ^+ denote the signature Σ without the symbol $\neg$, and For^+ be the corresponding $\neg$-free fragment of *For*. *Positive classical logic*, from now on denoted as $\mathbf{CPL}^+$, may be axiomatized in the signature Σ^+ by axioms (Ax1)–(Ax9), plus (MP). *Classical propositional logic*, from now on denoted by **CPL**, is an extension of $\mathbf{CPL}^+$ in the signature Σ, where $\neg$ is governed by two dual axioms, (Ax10) and the following 'explosion law':

(exp) $\alpha \rightarrow (\neg\alpha \rightarrow \beta)$

That axiomatization should come as no surprise, if you only recall the notion of a classical negation from Definition 8. Clearly, for any logic $\mathbf{L}$ extending $\mathbf{CPL}^+$ a (primitive or defined) unary connective $\neg$ of $\mathbf{L}$ is a classical negation iff the schemas $(\alpha \vee \neg\alpha)$ and $(\alpha \rightarrow (\neg\alpha \rightarrow \beta))$ are provable.

CPL is also the minimal consistent extension of C_1. Indeed, an alternative way of axiomatizing **CPL** is by adding $\circ\alpha$ to C_1 as a new axiom schema, and (exp) then follows from (bc1) and this new axiom, by (MP). On the other hand, *positive intuitionistic logic* may be axiomatized from $\mathbf{CPL}^+$ by dropping (Ax9).

As mentioned above, C_1 may be considered as a deductive fragment of **CPL** (in the signature Σ), whereas **CPL** may be considered as a deductive fragment of C_1 in the signature $\Sigma^{\sim}$ obtained from Σ by adding a symbol $\sim$ for classical negation, and where $\neg$ denotes the paraconsistent negation of C_1.

As it is well known (cf. [Mendelson, 1997]), any logic having (Ax1) and (Ax2) as axioms, and *modus ponens* (MP) as its only primitive inference rule has a deductive implication.[7]

In any logic endowed with a deductive implication, the Principle of Explosion, (3), and the explosion law, (exp), are interderivable. So, for any such logic, if paraconsistency is to be obtained, (exp) must fail.

As usual, bi-implication $\leftrightarrow$ will be defined here by setting $(\alpha \leftrightarrow \beta) \stackrel{\text{def}}{=} ((\alpha \rightarrow \beta) \wedge (\beta \rightarrow \alpha))$. Note that, in the presence of a deductive implication $\rightarrow$, $\vdash (\alpha \leftrightarrow \beta)$ if, and only if, $\alpha \vdash \beta$ and $\beta \vdash \alpha$, that is, iff α and β are equivalent. Nevertheless, the equivalence of two formulas, in the logics we will study here, does not necessarily guarantee that these formulas may be freely inter-substituted for each other, as we shall see below. ■

Recall that the definition of a **C**-system (Definition 27) mentioned **LFI**s in which the set $\bigcirc(p)$ could be taken as a singleton. The easiest way of realizing this intuition is by extending the original language of our logics so as to count from the start with a primitive connective $\circ$ for consistency.

REMARK 30. Recall the signature Σ° from Remark 15. Consider the following (innocuous, but linguistically relevant) extension of **CPL** that presupposes all formulas to be consistent, obtained by the addition of the following new axiom:

(ext) $\circ\alpha$

In practice, this will constitute of course just another version of **CPL** in a different signature, where any formula of the form $\circ\alpha$ is assumed to be a top particle. This logic, which we will here call *extended classical logic* and denote by **eCPL**, will come in handy below when we start building **C**-systems based on classical logic. ■

Sometimes our Logics of Formal Inconsistency can dismiss the new consistency connective (by replacing it by a formula built from the other connectives already present in the signature). Before defining this class of logics,

[7] This is not always true, though, for logics extending (Ax1), (Ax2) and (MP) by the addition of new primitive inference rules.

it is convenient to make a little detour and present a fundamental notion that will have a role to play in several parts of this chapter, namely, the concept of translation between logics.

DEFINITION 31. Let $\mathbf{L}2$ and $\mathbf{L}1$ be logics with sets of formulas For_2 and For_1, respectively. A mapping t: $For_2 \longrightarrow For_1$ is said to be a *translation from* $\mathbf{L}2$ *to* $\mathbf{L}1$ if, for every set $\Gamma \cup \{\alpha\}$ of $\mathbf{L}2$-formulas,

$$\Gamma \vdash_{\mathbf{L}2} \alpha \text{ implies } t(\Gamma) \vdash_{\mathbf{L}1} t(\alpha).$$

Here, $t(\Gamma)$ stands for $\{t(\gamma) : \gamma \in \Gamma\}$.

If 'implies' is replaced by 'iff' in the definition above, then t is called a *conservative translation*. See [da Silva *et al.*, 1999], [Coniglio and Carnielli, 2002] and [Coniglio, 2005] for a general account of translations and conservative translations. ■

Now, having the notion of translations at hand, the special kind of **C**-systems mentioned above is defined as follows:

DEFINITION 32. Let $\mathbf{L}2$ be a **C**-system with respect to $\neg$, based on a logic $\mathbf{L}1$, and let $\varphi(p)$ represent the formula schema with respect to which $\mathbf{L}2$ is gently explosive, that is, such that $\varphi(\alpha)$ represents in $\mathbf{L}2$ the consistency of the formula α with respect to the non-explosive negation $\neg$. Where Σ_2 represents the signature of the logic $\mathbf{L}2$, let $\mathsf{cnt}[\varphi(\mathsf{p})]$ represent the set of connectives involved in the formulation of $\varphi(p)$. Let Σ' be any signature obtained by dropping from Σ_2 all the connectives that appear in $\mathsf{cnt}[\varphi(\mathsf{p})]$, that is, Σ' is a restriction of the signature of $\mathbf{L}2$ in which consistency can no more be expressed in the same way as in the original logic $\mathbf{L}2$. Now, in case it is still possible to express the consistency of the formulas of $\mathbf{L}2$ with the help of the remaining connectives in $\Sigma' \subsetneq \Sigma_2$, say, by way of a set of formulas $\varphi'(p)$ over Σ', then we say that $\mathbf{L}2$ is a **dC**-*system* based on $\mathbf{L}1$ (or simply a **dC**-*system*). So, **dC**-systems are **C**-systems with respect to some negation and some consistency schema $\varphi(p)$ where it is also possible to express consistency alternatively by way of a formula $\varphi'(p)$ such that $\varphi(p)$ and $\varphi'(p)$ have no common structure, that is, such that $\mathsf{cnt}[\varphi'(\mathsf{p})] \cap \mathsf{cnt}[\varphi'(\mathsf{p})] = \varnothing$. This is typically the case when $\varphi(p)$ has the form $\circ(p)$, where $\circ$ is a primitive unary connective of Σ_2, but where, at the same time, $\circ$ can be explicitly defined by way of the connectives in $\Sigma_2 \setminus \{\circ\}$ (see examples below). In that case we say that $\mathbf{L}2$ is a *direct* **dC**-*system* based on $\mathbf{L}1$ (or simply a *direct* **dC**-*system*). As we will see below, there are **dC**-systems that are not direct (they will from here on be called *indirect*). In those indirect **dC**-systems, consistency cannot be expressed by a unary connective $\circ$, primitive or defined, but only by way of a complex formula φ, depending on a single variable.

DEFINITION 33. Let Σ be the signature of an indirect **dC**-system $\mathbf{L}$, and consider the direct **dC**-system $\mathbf{L}'$ defined over the signature Σ', such that:

(a) $\mathbf{L}'$ is a conservative extension of $\mathbf{L}$ obtained by the addition of a new unary connective $\circ$, that is, such that $\Sigma' = \Sigma \cup \{\circ\}$ (so, in particular, the consistency of a formula α can be expressed in $\mathbf{L}'$ exactly as in $\mathbf{L}$, namely, by way of the formula $\varphi(\alpha)$);

(b) $\mathbf{L}'$ is an **LFI** with respect to $\circ \in \Sigma'$, and a **C**-system with respect to some $\neg \in \Sigma$ (so, in particular, the consistency of α can also be expressed in $\mathbf{L}'$ by way of the formula $\circ\alpha$);

(c) In $\mathbf{L}'$ the connective $\circ$ plays the same role as the formula φ plays in $\mathbf{L}$, more specifically, there is a translation $\star : For_{\mathbf{L}'} \longrightarrow For_{\mathbf{L}}$ respecting the following clauses:

(c.1) $t(p) = p$, for p a propositional variable

(c.2) $t(\divideontimes(\alpha_1, \ldots, \alpha_n)) = \divideontimes(t(\alpha_1), \ldots, t(\alpha_n))$, for every n-ary connective $\divideontimes$ in Σ' distinct from $\circ$, and for any choice of formulas $\alpha_1, \ldots, \alpha_n$ from $For_{\mathbf{L}'}$

(c.3) $t(\circ\alpha) = \varphi(t(\alpha))$, for any formula α in $For_{\mathbf{L}'}$

In a case like this we may say that the direct **dC**-system $\mathbf{L}'$ *corresponds* to the indirect **dC**-system $\mathbf{L}$. Indirect **dC**-systems appear typically when we are talking about **C**-systems for which the replacement property fails to such an extent that it might turn out to be impossible to give an explicit definition of the consistency connective in terms of other, more usual connectives. (Examples follow below.) ■

The next example and the subsequent theorem will show that **dC**-systems are even more ubiquitous than one might initially imagine.

EXAMPLE 34. Let $\Sigma^{\Diamond\Box}$ be the signature obtained by the addition of the new unary connectives $\Diamond$ and $\Box$ to the signature Σ, where the connectives $\wedge$, $\vee$, $\rightarrow$ and $\neg$ of Σ are interpreted as in classical logic and the new connectives are interpreted as usual in normal modal logics. So, $\Diamond\alpha$ (respectively, $\Box\alpha$) will be true in a given world iff α is true in some (respectively, any) world accessible to the former. The most obvious degenerate examples of normal modal logics are characterized by frames that are such that every world can access only itself or no other world. As shown in [Marcos, 2005e], inside any non-degenerate normal modal logic, a paraconsistent negation $\backsim$ may be defined by setting $\backsim\alpha \stackrel{\text{def}}{=} \Diamond\neg\alpha$, and a consistency connective may be defined by setting $\circ\alpha \stackrel{\text{def}}{=} \alpha \rightarrow \Box\alpha$.

Conversely, take the signature Σ°, and interpret the primitive negation $\neg$ now over Kripke structures so as to make it behave exactly like the above connective $\backsim$, that is, an interpretation such that, for worlds x and y of a model $\mathcal{M}$ with an accessibility relation R:

$$\models_x^{\mathcal{M}} \neg\alpha \text{ iff } (\exists y)(x\mathsf{R}y \text{ and } \not\models_y^{\mathcal{M}} \alpha).$$

Moreover, let the consistency connective be interpreted in such a way that:

$$\models_x^{\mathcal{M}} \circ\alpha \text{ iff } \models_x^{\mathcal{M}} \alpha \text{ implies } (\forall y)(\text{if } x\mathsf{R}y \text{ then } \models_y^{\mathcal{M}} \alpha).$$

Then, in the present case, one can still redefine the previous connectives of $\Sigma^{\Diamond\Box}$. Indeed, one can define a bottom $\bot$ by setting $\bot \stackrel{\text{def}}{=} \alpha \wedge (\neg\alpha \wedge \circ\alpha)$, for an arbitrary formula α, and then define a classical negation $\sim$ by setting $\sim\alpha \stackrel{\text{def}}{=} \alpha \rightarrow \bot$. The original modal connectives can finally be defined by setting $\Diamond\alpha \stackrel{\text{def}}{=} \neg\sim\alpha$ and $\Box\alpha \stackrel{\text{def}}{=} \sim\neg\alpha$.

The above arguments show that any non-degenerate normal modal logic may be naturally reformulated in the signature of an **LFI**. In that sense, modal logics are typically paraconsistent, and could be recast as the study of paraconsistent negations (instead of operators such as $\Box$ and $\Diamond$). ■

THEOREM 35.
(i) **LFI1** (see Example 18) is a **C**-system (based either on $\mathbf{CPL}^+$ or on **CPL**), but not a **dC**-system.

(ii) $\mathbf{P}^1$ (see Example 19) is a direct **dC**-system.

(iii) The logics $C_n, 1 \leq n < \omega$, (see Definition 28) are all direct **dC**-systems.

(iv) Jaśkowski's Discussive Logic **D2** (see Example 24) is a direct **dC**-system.

(v) The normal modal logics from Example 34 are all direct **dC**-systems.

Proof. For item (i), observe first that **LFI1** is a **C**-system based on classical logic. Indeed, the binary connectives of **LFI1** all behave classically: All axioms of $\mathbf{CPL}^+$ are validated by the 3-valued truth-tables of **LFI1**, and (MP) preserves validity. Second, as we already know, the classical negation $\sim$ can be defined in **LFI1**. Third, the connective $\circ$ expresses consistency in **LFI1**, and the latter logic is indeed a conservative extension of *Pac* obtained exactly by the addition of that connective. Similarly, the non-explosive negation $\neg$ of **LFI1** can easily be seen not to be definable, in **LFI1**, from the truth-tables of the classical connectives. Finally, recall from Theorem 25 that *Pac* is not an **LFI**, and observe that $\circ$ is not definable from the other connectives of **LFI1**. Items (ii)–(v) were already explained when the corresponding logics were introduced. ■

The first examples of indirect **dC**-systems will appear only in Theorems 106 and 110, as well as Remark 111.

All **LFI**s studied from the next subsection on, unless explicit mention to the contrary, are **C**-systems based on classical logic, and can therefore be axiomatized starting from $\mathbf{CPL}^+$.

*3.2 Towards **mbC**, a fundamental **LFI***

Before introducing our weakest **LFI** based on classical logic, we will introduce a very weak non-gently explosive paraconsistent logic.

Do bear in mind, from Remark 29, that $\neg$ in **CPL** was axiomatized by the addition to $\mathbf{CPL}^+$ of two dual clauses, (Ax10) and (exp).

DEFINITION 36. The paraconsistent logic *PI*, investigated in [Batens, 1980], extends $\mathbf{CPL}^+$ in the signature Σ (see Remark 29) by the addition of (Ax10). In other words, *PI* is axiomatized by (Ax1)–(Ax10) and (MP) (recall Definition 28). ■

It is worth noting that, due to (Ax8), (Ax10) and to the fact that *PI* has a deductive implication (recall Definition 6), one can count on the classical proof strategy known as *proof-by-cases*:

THEOREM 37. If $(\Gamma, \alpha \vdash_{PI} \beta)$ and $(\Delta, \neg\alpha \vdash_{PI} \beta)$ then $(\Gamma, \Delta \vdash_{PI} \beta)$. ■

Here are some other important properties of *PI*:

THEOREM 38. (i) *PI* is boldly paraconsistent.

Moreover, for any boldly paraconsistent extension $\mathbf{L}$ of *PI*:

(ii) *Reductio ad absurdum* is not a valid rule, i.e. rules such as:
$(\Delta, \beta \vdash_{\mathbf{L}} \alpha)$ and $(\Pi, \beta \vdash_{\mathbf{L}} \neg\alpha)$ implies $(\Delta, \Pi \vdash_{\mathbf{L}} \neg\beta)$, and
$(\Delta, \neg\beta \vdash_{\mathbf{L}} \alpha)$ and $(\Pi, \neg\beta \vdash_{\mathbf{L}} \neg\alpha)$ implies $(\Delta, \Pi \vdash_{\mathbf{L}} \beta)$
cannot obtain.

(iii) If the implication $\rightarrow$ is still a deductive implication (recall Definition 6), contraposition is not a valid rule, i.e. rules such as:
$\Gamma, \alpha \rightarrow \beta \vdash_{\mathbf{L}} \neg\beta \rightarrow \neg\alpha$
$\Gamma, \alpha \rightarrow \neg\beta \vdash_{\mathbf{L}} \beta \rightarrow \neg\alpha$
$\Gamma, \neg\alpha \rightarrow \beta \vdash_{\mathbf{L}} \neg\beta \rightarrow \alpha$
$\Gamma, \neg\alpha \rightarrow \neg\beta \vdash_{\mathbf{L}} \beta \rightarrow \alpha$
cannot obtain.

Proof. For item (i), note that *PI* has a deductive implication and is a fragment of both *Pac* and $\mathbf{P}^1$. Indeed, the axioms of *PI* are all validated by the truth-tables of *Pac* and by the truth-tables of $\mathbf{P}^1$, and (MP) preserves validity. Recall that those 3-valued extensions of *PI* were already proven to be boldly paraconsistent in Theorem 20.
For item (ii), let $\Delta = \Pi = \{\alpha, \neg\alpha\}$. Then, by *reductio*, the logic would be partially explosive.
For item (iii), using the properties of the deductive implication, we have that $\gamma \vdash_{\mathbf{L}} \alpha \rightarrow \gamma$. Then again, by contraposition, the logic would turn out to be partially explosive. ■

As we will soon see (check Theorem 48), the upgrade of non-gently explosive logics into **LFI**s will help remedy the above mentioned deductive weaknesses, so typical of paraconsistent logics in general.

Here again, using the fact that *PI* is a deductive fragment of *Pac*, it can also be easily checked that:

THEOREM 39. The logic *PI*:
(i) does not have a supplementing negation, nor a bottom particle;
(ii) is not finitely trivializable;
(iii) is not an **LFI**. ■

Before proceeding, this seems to be a convenient place to mention some logics that live very close to *PI*:

DEFINITION 40. The logic C_{min} (cf. [Carnielli and Marcos, 1999]) is obtained from *PI* by the addition of $\neg\neg\alpha \rightarrow \alpha$ as a new axiom. The logic C_ω (cf. [da Costa, 1963]) is obtained from C_{min} by dropping (Ax9). Finally, the logic *CAR* (cf. [da Costa and Béziau, 1993]) is obtained from *PI* by adding $\alpha \rightarrow (\neg\alpha \rightarrow \neg\beta)$ as a new axiom. ■

Finally, here are some other important facts about *PI*:

THEOREM 41.
(i) *PI* does not prove any negated formula (that is, any formula of the form $\neg\delta$).
(ii) No two different negated formulas of *PI* are equivalent, that is, if $\neg\alpha \dashv\vdash_{PI} \neg\beta$ then $\alpha = \beta$.

Proof. Item (i) was already proven in [Carnielli and Marcos, 1999] for C_{min}. Item (ii) was proven in [Urbas, 1989] for C_ω, and the proof may be easily adapted for *PI*. ■

As we saw in Theorem 39(iii), *PI* is not an **LFI**. We will now make the most obvious upgrade of *PI* that will turn it into an **LFI**, endowing it with the most straightforward axiomatic version of the principle (10), the so-called Finite Gentle Principle of Explosion:

DEFINITION 42. Recall the signature Σ° from Remark 15 and the logic *PI* from Definition 36. The logic **mbC** is obtained from *PI*, over Σ°, by the addition of the following axiom schema:

(bc1) $\circ\alpha \rightarrow (\alpha \rightarrow (\neg\alpha \rightarrow \beta))$

In other words, **mbC** is axiomatized by (Ax1)–(Ax10) plus (MP) (recall Definition 28), but now over the signature Σ°, together with the extra axiom (bc1), above. ■

Notice that a particular form of axiom (bc1) had already been considered in Definition 28, but there $\circ\alpha$ was considered as an abbreviation for $\neg(\alpha \wedge \neg\alpha)$, instead of a primitive connective. We recall that the intended reading of $\circ\alpha$ is 'α is consistent'. As we shall see, in general, $\circ\alpha$ is logically independent from $\neg(\alpha \wedge \neg\alpha)$.

If $\vdash_{\mathbf{mbC}}$ denotes the consequence relation of **mbC**, then we obtain, by (MP), the following:

$$\circ\alpha,\ \alpha,\ \neg\alpha \vdash_{\mathbf{mbC}} \beta \tag{11}$$

Rule (11) may be read as saying that 'if α is consistent and contradictory, then it explodes'. Clearly, this rule amounts to a realization of the Finite Gentle Principle of Explosion (10), as in our formulation of da Costa's C_n (Definition 28), with the difference that now $\circ$ is a primitive unary connective and *not* an abbreviation depending on conjunction and negation.

REMARK 43. It is easy to define supplementing negations in **mbC**. Consider first a negation $\wr$ set by $\wr\alpha \stackrel{\text{def}}{=} (\neg\alpha \wedge \circ\alpha)$. Notice that, as a particular instance of Theorem 13(i), $\bot_\beta \stackrel{\text{def}}{=} (\beta \wedge \wr\beta)$ defines a bottom particle, for every β. Consider next a negation $\sim_\beta$ set by $\sim_\beta\alpha \stackrel{\text{def}}{=} \alpha \to \bot_\beta$. Clearly, $\forall\alpha\forall\gamma(\alpha, \wr\alpha \vdash_{\mathbf{mbC}} \gamma)$ and $\forall\beta\forall\alpha\forall\gamma(\alpha, \sim_\beta\alpha \vdash_{\mathbf{mbC}} \gamma)$. In Remark 70, the semantic tools of Subsection 3.4, granting sound and complete possible-translations interpretations for **mbC**, will help us showing that neither $\sim_\beta\alpha$ nor $\wr\alpha$ are always bottom particles. Moreover, these supplementing negations will in fact be seen to be inequivalent: though $\wr\alpha$ derives $\sim_\beta\alpha$, the converse is not true. While $\sim_\beta$ defines a classical negation, $\wr$ fails to be complementing (the latter facts will be proven in Remark 70).

From now on, we will simply write $\bot$ and $\sim$ to refer to any of the connectives $\bot_\beta$ and $\sim_\beta$ defined above. Despite $\bot_\beta$ and $\bot_\gamma$, as well as $\sim_\beta\alpha$ and $\sim_\gamma\alpha$, being equivalent for every β, γ and α, they cannot be freely intersubstituted (check the end of Remark 29). It will be often useful, in this paper, to consider our **C**-systems to be written from the start in an extended signature containing both the non-explosive negation $\neg$ and the classical negation $\sim$, to be set as in the above definition. ■

THEOREM 44. **mbC** is an **LFI**. In fact, it is a **C**-system based on **CPL**.

Proof. Note that **mbC** is indeed a fragment of **LFI1** and of $\mathbf{P}^1$, and in Theorem 25 the latter were shown to be **LFI**s. Moreover, we now know from rule (11) that the principle (9) holds in **mbC** (in fact its finite form (10) already holds). By design, we also know that **mbC** contains $\mathbf{CPL}^+$, and $\neg$ cannot be defined in the latter logic. Thus, **mbC** is a **C**-system based on $\mathbf{CPL}^+$ such that $\bigcirc(p) = \{\circ p\}$. To check that **mbC** can also be seen as a **C**-system based on full **CPL** one might notice that **mbC** extends **CPL** in a signature with two negations (as in the preceding remark). This extension must be conservative, given that **CPL** is well-known to be maximal with respect to the trivial logic. ■

So, **mbC** may be considered as a deductive fragment of **CPL**, provided that **CPL** is presented as **eCPL** in the signature Σ°. On the other hand, taking into account the signature $\Sigma^{\circ\sim}$ obtained from Σ° by adding a symbol $\sim$ for the classical negation $\sim\alpha = \alpha \to \bot$ of **mbC** (recall Remark 43), and where $\neg$ denotes the paraconsistent negation, **CPL** is a deductive fragment of **mbC** such that **mbC** is a **C**-system based on **CPL**, provided that the obvious axioms defining $\sim$ in terms of the other connectives of Σ° are added to **mbC**.

REMARK 45. In spite of the term 'Logics of Formal *In*consistency', we have mentioned but a *consistency* connective $\circ$ this far. But **mbC** can also count on the dual *inconsistency* connective $\bullet$. To define it, in general, one might make use of a classical negation, such as the negation $\sim$ defined in the above remark, and set $\bullet\alpha \stackrel{\text{def}}{=} \sim\circ\alpha$. ■

The logic **mbC** inherits the main properties of the positive fragment of *PI* (such as those properties of the standard conjunction, the standard disjunction and the deductive implication), but above we have seen that the former logic is much richer than the latter. As another illustration of this fact, from Theorem 44 and Remark 43 we can immediately see that none of the claims from Theorem 39 are any longer valid in **mbC**. Furthermore, the claims of Theorem 41 also do not hold good for **mbC**:

THEOREM 46.
(i) There are in **mbC** theorems of the form $\neg\delta$, for some formula δ.
(ii) There are formulas α and β in **mbC** such that $\alpha \neq \beta$, α and β are equivalent, and $\neg\alpha$ and $\neg\beta$ are also equivalent.

Proof. (i) Consider any bottom particle $\bot$ of **mbC**. Then $(\bot \vdash_{\mathbf{mbC}} \neg\bot)$ and $(\neg\bot \vdash_{\mathbf{mbC}} \neg\bot)$, thus $\vdash_{\mathbf{mbC}} \neg\bot$, by Theorem 37.
(ii) Take α and β to be any two syntactically distinct bottom particles. ■

Even if, differently from *PI*, **mbC** *does* have negated theorems, it does *not* have consistent theorems:

THEOREM 47. There are in **mbC** no theorems of the form $\circ\delta$.

Proof. Use the classical truth-tables over $\{0, 1\}$ for $\wedge, \vee, \rightarrow$ and $\neg$, and pick for $\circ$ a truth-table with value constant and equal to 0. ■

The price to pay for paraconsistency is that we necessarily lose some theorems and inferences dependent on the 'consistency presupposition'. This has been illustrated, for instance, in Theorem 38, where *PI* and its extensions (satisfying certain assumptions) were shown to lack some usual classical proof strategies such as *reductio* and contraposition. This loss in inferential power can be remedied in the **LFI**s exactly by adding convenient consistency assumptions at the object-language level, as advanced in Remark 26. Indeed, some restricted forms of those rules may be proven in **mbC**:

THEOREM 48. The following *reductio* rules hold good in **mbC**:

(i) $(\Gamma \vdash_{\mathbf{mbC}} \circ\alpha)$ and $(\Delta, \beta \vdash_{\mathbf{mbC}} \alpha)$ and $(\Lambda, \beta \vdash_{\mathbf{mbC}} \neg\alpha)$ implies $(\Gamma, \Delta, \Lambda \vdash_{\mathbf{mbC}} \neg\beta)$

(ii) $(\Gamma \vdash_{\mathbf{mbC}} \circ\alpha)$ and $(\Delta, \neg\beta \vdash_{\mathbf{mbC}} \alpha)$ and $(\Lambda, \neg\beta \vdash_{\mathbf{mbC}} \neg\alpha)$ implies $(\Gamma, \Delta, \Lambda \vdash_{\mathbf{mbC}} \beta)$

The following contraposition rules hold in **mbC**:

(iii) $\circ\beta, (\alpha \rightarrow \beta) \vdash_{\mathbf{mbC}} (\neg\beta \rightarrow \neg\alpha)$

(iv) $\circ\beta, (\alpha \rightarrow \neg\beta) \vdash_{\mathbf{mbC}} (\beta \rightarrow \neg\alpha)$

(v) $\circ\beta, (\neg\alpha \rightarrow \beta) \vdash_{\mathbf{mbC}} (\neg\beta \rightarrow \alpha)$

(vi) $\circ\beta, (\neg\alpha \rightarrow \neg\beta) \vdash_{\mathbf{mbC}} (\beta \rightarrow \alpha)$ ■

The last theorem is an instance of a more general phenomenon: Any classical rule may be recovered within our **C**-systems based on classical logic (check the discussion about that at Subsection 3.6).

Intuitively, a contradiction might be seen as a sufficient condition for inconsistency. Indeed, here are some properties that relate the new connective of consistency to the more familiar connectives of $\mathbf{CPL}^+$:

THEOREM 49. In **mbC** the following hold good:

(i) $\alpha, \neg\alpha \vdash_{\mathbf{mbC}} \neg\circ\alpha$

(ii) $\alpha \wedge \neg\alpha \vdash_{\mathbf{mbC}} \neg\circ\alpha$

(iii) $\circ\alpha \vdash_{\mathbf{mbC}} \neg(\alpha \wedge \neg\alpha)$

(iv) $\circ\alpha \vdash_{\mathbf{mbC}} \neg(\neg\alpha \wedge \alpha)$

The converses of these rules are all failed by **mbC**.

Proof. Items (i)–(iv) are easy consequences of the restricted forms of *reductio* from Theorem 48.

In order to prove the second half of the theorem, consider the truth-tables of $\mathbf{P}^1$ (Example 19), but substitute the truth-table for negation, $\neg$, by the 3-valued truth-table for classical negation, $\sim$, to be found in Example 17. Then, **mbC** is sound for this set of truth-tables, and so it is enough to prove the failure of the converse rules using these same truth-tables. For instance, the rule $\neg(\neg\alpha \wedge \alpha) \vdash \circ\alpha$, converse to rule (iv), is failed if we put an atom p in the place of the schema α and pick a valuation v such that $v(p) = \frac{1}{2}$. Indeed, observe that the above described set of truth-tables will make $v(\neg p) = 0$, thus $v(p \wedge \neg p) = 0$ and $v(\neg(p \wedge \neg p)) = 1$, while they will also make $v(\circ p) = 0$, providing a counter-model for this inference that is nevertheless sound for **mbC**. (Alternative counter-models, in terms of possible-translations semantics, will be offered in Example 69.) ■

The last result hints to the fact that paraconsistent logics may easily have certain unexpected asymmetries. That's what happens, for instance, with da Costa's C_1. As we shall see, the converse of (iii) holds good in C_1, while the converse of (iv) fails, so that $\neg(\alpha\wedge\neg\alpha)$ and $\neg(\neg\alpha\wedge\alpha)$ are not equivalent formulas in C_1. Other even more shocking examples of asymmetries are the following:

THEOREM 50. In **mbC**:

(i) $(\alpha \wedge \beta) \dashv\vdash_{\mathbf{mbC}} (\beta \wedge \alpha)$ holds good,
but $\neg(\alpha \wedge \beta) \dashv\vdash_{\mathbf{mbC}} \neg(\beta \wedge \alpha)$ does not hold.

(ii) $(\alpha \vee \beta) \dashv\vdash_{\mathbf{mbC}} (\beta \vee \alpha)$ holds good,
but $\neg(\alpha \vee \beta) \dashv\vdash_{\mathbf{mbC}} \neg(\beta \vee \alpha)$ does not hold.

(iii) $(\alpha \wedge \neg\alpha) \dashv\vdash_{\mathbf{mbC}} (\neg\alpha \wedge \alpha)$ holds good,
but $\neg(\alpha \wedge \neg\alpha) \dashv\vdash_{\mathbf{mbC}} \neg(\neg\alpha \wedge \alpha)$ does not hold.

(iv) $\gamma \vee \neg\gamma$ is a top particle, thus $(\alpha \vee \neg\alpha) \dashv\vdash_{\mathbf{mbC}} (\beta \vee \neg\beta)$ holds good. But $\neg(\alpha \vee \neg\alpha) \dashv\vdash_{\mathbf{mbC}} \neg(\beta \vee \neg\beta)$ does not hold.

(v) The equivalence $\alpha \dashv\vdash_{PI} (\neg\alpha \rightarrow \alpha)$ holds good, but $\neg\alpha \dashv\vdash_{\mathbf{mbC}} \neg(\neg\alpha \rightarrow \alpha)$ does not hold.

Proof. Using *PI* it is easy to prove the first halves of each item.

Items (i) to (iii). In order to check that none of the other halves hold, we can use again the truth-tables of $\mathbf{P}^1$ (Example 19), but redefining $(1 \wedge \frac{1}{2}) = (1 \vee \frac{1}{2}) = \frac{1}{2}$.

For item (iv), use the truth-tables of **LFI1** (Example 18), and take a valuation v such that $v(p) \neq v(q)$ and $v(p), v(q) \in \{1, \frac{1}{2}\}$. For item (v), use again the truth-tables of $\mathbf{P}^1$, and consider $v(p) = \frac{1}{2}$. ■

REMARK 51. The last theorem illustrates the failure of the so-called *replacement property*. This property states that, for any choice of formulas $\alpha_0, \ldots, \alpha_n$, $\beta_0, \ldots, \beta_n$ and of formula $\varphi(p_0, \ldots, p_n)$:

(RP) $(\alpha_0 \dashv\Vdash \beta_0)$ and $\ldots$ and $(\alpha_n \dashv\Vdash \beta_n)$ implies $\varphi(\alpha_0, \ldots, \alpha_n) \dashv\Vdash \varphi(\beta_0, \ldots, \beta_n)$

For example, from $\alpha \dashv\Vdash \beta$ one would immediately derive $\neg\alpha \dashv\Vdash \neg\beta$, using (RP). But this does not hold for **mbC**. Recall, by the way, that $\alpha \dashv\vdash_{\mathbf{mbC}} \beta$ amounts to $\vdash_{\mathbf{mbC}} \alpha \leftrightarrow \beta$, given the definition of bi-implication and the presence of a deductive implication in **mbC**. Logics enjoying (RP) are called *self-extensional* in [Wójcicki, 1988]. Paradigmatic examples of such logics are provided by normal modal logics. ■

We will show below that various other classes of **LFI**s fail the replacement property (see Theorems 52, 81 and 133).

A natural question here is whether our logics can be upgraded so as to restore the interesting property (RP) inside the paraconsistent territory. To ensure that (RP) is obtainable in extensions of *PI* in the signature Σ, it is enough to add the rule:

(EC) $\forall\alpha\forall\beta((\alpha \dashv\Vdash \beta)$ implies $(\neg\alpha \dashv\Vdash \neg\beta))$

In [Urbas, 1989] paraconsistent extensions of C_ω (see Definition 40) enjoying the rule (EC) are shown to exist. The argument may be easily adapted to several extensions of *PI*, but it does not follow for many other such extensions, as it will be shown below. In [da Costa and Béziau, 1993], the logic *CAR* (see Definition 40) was introduced as an extension of *PI* where (RP) holds good. But *CAR* is not an **LFI**, and it is not boldly paraconsistent, being partially explosive exactly as the Minimal Intuitionistic Logic *MIL* from Example 10. To obtain the replacement property in extensions of **mbC**, in the signature Σ°, a further rule is needed, namely:

(EO) $\forall\alpha\forall\beta((\alpha \dashv\Vdash \beta)$ implies $(\circ\alpha \dashv\Vdash \circ\beta))$

Before ending this subsection, let us quickly survey some results on the possible validity of (RP) in paraconsistent extensions of **mbC**, or in some of its fragments:

THEOREM 52. The replacement property (RP) cannot hold in any paraconsistent extension of **mbC** in which:

(i) $\circ\neg\neg\alpha$ holds, for some given classical negation $\neg$; or

(ii) $\neg(\alpha \rightarrow \beta) \vdash (\alpha \wedge \neg\beta)$ holds.

The replacement property (RP) cannot hold in any left-adjunctive paraconsistent extension of *PI* in which:

(iii) $(\alpha \wedge \beta) \dashv\vdash \neg(\neg\alpha \vee \neg\beta)$ holds.

The replacement property (RP) cannot hold in any left-adjunctive paraconsistent logic in which:

(iv) $\neg(\alpha \wedge \neg\alpha)$ holds and $(\alpha \wedge \neg\alpha) \dashv\vdash \neg\neg(\alpha \wedge \neg\alpha)$.

Proof. Assume that (i) holds good. Since $\neg$ is a classical negation, $\alpha \dashv\vdash \neg\neg\alpha$ and then, by (RP), we infer that $\circ\alpha \dashv\vdash \circ\neg\neg\alpha$. But $\circ\neg\neg\alpha$ is a theorem of the given logic, by hypothesis, then $\circ\alpha$ is a theorem. From (bc1), the logic turns out to be explosive with respect to the original primitive negation $\neg$. Now, assume that (ii) holds good. Consider the supplementing negation $\sim\alpha = (\alpha \rightarrow \bot)$ for **mbC**, where $\bot = (p_0 \wedge (\neg p_0 \wedge \circ p_0))$, proposed in Remark 43. This negation was shown to be classical. Then, $\neg\sim\alpha \vdash (\alpha \wedge \neg\bot)$, by hypothesis, and so $\neg\sim\alpha \vdash \alpha$, using (Ax4). Since $\alpha, \sim\alpha \vdash \beta$ for every α and β, then $\neg\sim\alpha, \sim\alpha \vdash \beta$ for every α and β, that is, the logic is controllably explosive in contact with $\sim p$. In particular, $\neg\sim\sim\alpha, \sim\sim\alpha \vdash \beta$ for every α and β. But $\alpha \dashv\vdash \sim\sim\alpha$ for a classical negation and so, using (RP), we may conclude that $\neg\alpha \dashv\vdash \neg\sim\sim\alpha$ and then $\neg\alpha, \alpha \vdash \beta$ for every β. In other words, the logic will be explosive, not paraconsistent (with respect to the original negation $\neg$).

Assume next that (iii) holds good. Since $(\neg\alpha \vee \neg\neg\alpha)$ is a theorem of *PI*, then $\neg(\neg\alpha \vee \neg\neg\alpha) \dashv\vdash \neg(\neg\beta \vee \neg\neg\beta)$, for every α and β, by (RP). By hypothesis we infer that $(\alpha \wedge \neg\alpha) \dashv\vdash (\beta \wedge \neg\beta)$. So, by the rules of a standard conjunction, we conclude in particular that $\alpha, \neg\alpha \vdash \beta$.

Finally, assume that (iv) holds good. Since $\neg(\alpha \wedge \neg\alpha)$ is a theorem, by hypothesis, then $\neg\neg(\alpha \wedge \neg\alpha) \dashv\vdash \neg\neg(\beta \wedge \neg\beta)$ for every α and β, by (RP). Then, again by hypothesis, we have that $(\alpha \wedge \neg\alpha) \dashv\vdash (\beta \wedge \neg\beta)$. The result follows now as in item (iii). ■

With the help of Theorem 52(ii) it is easy to see, for instance, that Jaśkowski's **D2** (recall Example 24) fails the replacement property. This feature was used in [Marcos, 2005b] to show that this logic is not 'modal' in the current usual sense of the word, in spite of its very definition in terms of a double translation into the modal logic $S5$.

REMARK 53. To obtain paraconsistent extensions of **mbC** validating both (EC) and (EO) is a perfectly feasible task. Examples of such logics were

already offered in Example 34: Notice indeed that axiom (bc1) and rules (EC) and (EO) are all satisfied by the minimal normal modal logic K, thus also by any of its normal modal extensions. ■

3.3 Bivaluation semantics for ***mbC***

At the beginning of their historical trajectory, most **C**-systems were introduced exclusively in proof-theoretical terms (see, for a survey, [Carnielli and Marcos, 2002]). Later on, many of them were proven not to be characterizable by finite-valued truth-tables (such results are generalized here in Theorems 121 and 125). If we add to this the frequent failure of the replacement property and the consequent difficulty in characterizing those same logics by way of usual Kripke-like modal semantics, it will seem clear that semantic presentations for many of our present **C**-systems will have to rely upon some alternative kinds of semantics.

There are of course many examples of paraconsistent logics with adequate *finite-valued semantics*. Several 3-valued samples of such logics were already mentioned above in Examples 17, 18 and 19), and many more will be presented below in Section 5.3. Additionally, many examples of paraconsistent logics with a *modal semantics* were also mentioned above, in Example 34. However, we have already seen that a logic such as **mbC**, our weakest **LFI** based on classical logic, fails the replacement property. Moreover, as a particular consequence of Theorem 121, **mbC** will also be seen not to be finite-valued. What kind of semantics can we attach to such a logic, thus?

The first examples of adequate *non-truth-functional bivalued semantics* were proposed in [da Costa and Alves, 1977] in order to provide interpretations for some historically distinguished **C**-systems, those in the hierarchy C_n, $1 \leq n < \omega$ (check Definition 28). Such decidable semantics are now known to be a particular case of a more general semantic presentation, called 'dyadic' (check Subsection 3.5 and [Caleiro *et al.*, 2005a]). We will show in the following how a simple characteristic (non-truth-functional) adequate bivaluation semantics may be attached to the logic **mbC**. This example will help in clarifying the connections with other semantic presentations, as well as in devising relevant open problems towards obtaining a theoretical framework for further investigation in the foundations of paraconsistent logic. In the next subsection, we will endow **mbC** with the much richer semantics of possible-translations. This new semantics, as we shall see, not only gives an interpretation to contradictory situations, but it also offers an explanation for the existence of conflicting scenarios.

DEFINITION 54. Let $\mathbf{2} \stackrel{\text{def}}{=} \{0, 1\}$ be the set of truth-values, where 1 denotes the 'true' value and 0 denotes the 'false' value. An **mbC**-valuation is any function v: $For^{\circ} \longrightarrow \mathbf{2}$ subject to the following clauses:

(v1) $v(\alpha \wedge \beta) = 1$ iff $v(\alpha) = 1$ and $v(\beta) = 1$;

($v2$) $v(\alpha \vee \beta) = 1$ iff $v(\alpha) = 1$ or $v(\beta) = 1$;
($v3$) $v(\alpha \rightarrow \beta) = 1$ iff $v(\alpha) = 0$ or $v(\beta) = 1$;
($v4$) $v(\neg\alpha) = 0$ implies $v(\alpha) = 1$;
($v5$) $v(\circ\alpha) = 1$ implies $v(\alpha) = 0$ or $v(\neg\alpha) = 0$. ■

For a collection $\Gamma \cup \{\alpha\}$ of formulas of **mbC**, $\Gamma \vDash_{\mathbf{mbC}} \alpha$ means, as usual (recall Definition 16), that α is assigned the value 1 for every **mbC**-valuation that assigns value 1 to the elements of Γ.

REMARK 55. Given clause ($v5$) in the above definition of a bivaluation semantics for **mbC**, it is clear that this logic does not admit of a trivial model, that is, that there is no v such that $v(\alpha) = 1$ for every formula α. In particular, given a trivial theory Γ of **mbC**, for every **mbC**-valuation v, then there must be some $\gamma \in \Gamma$ such that $v(\gamma) = 0$ (and thus $v(\neg\gamma) = 1$, by clause ($v4$)). This observation reveals a typical semantical feature of **LFI**s. Indeed, other non-gently explosive paraconsistent logics might well allow for such trivial models. For instance, the logic *Pac* (Example 17), despite being maximal relative to classical logic (cf. [Batens, 1980]), does admit of such a model: Consider indeed $v(\alpha) = \frac{1}{2}$, and recall that $\frac{1}{2}$ is a designated value. ■

The soundness proof for **mbC** with respect to **mbC**-valuations is immediate:

THEOREM 56. [Soundness] Let $\Gamma \cup \{\alpha\}$ be a set of formulas in For°. Then: $\Gamma \vdash_{\mathbf{mbC}} \alpha$ implies $\Gamma \vDash_{\mathbf{mbC}} \alpha$.

Proof. Just check that all axioms of **mbC** assume only the value 1 in any **mbC**-valuation, and that (MP) preserves validity. ■

In order to prove completeness it is convenient to prove first some auxiliary lemmas. Let $\Delta \cup \{\alpha\}$ be a set of formulas in For°. We say that a theory Δ is *relatively maximal with respect to* α *in* **mbC** if $\Delta \nvdash_{\mathbf{mbC}} \alpha$ and for any formula β in For° such that $\beta \notin \Delta$ we have $\Delta, \beta \vdash_{\mathbf{mbC}} \alpha$. The usual Lindenbaum-Asser argument (cf. [Béziau, 1999]) shows that inside any compact S-logic — such as **mbC** — every non-trivial theory may be extended into a relatively maximal theory:

LEMMA 57. Let $\mathbf{L}$ be a compact S-logic over a signature $\widehat{\Sigma}$. Given some set of formulas Γ and a formula α such that $\Gamma \nvdash_{\mathbf{L}} \alpha$, then there is a set $\Delta \supseteq \Gamma$ that is relatively maximal with respect to α in $\mathbf{L}$.

Proof. Consider an enumeration $\{\varphi_n\}_{n\in\mathbb{N}}$ of the formulas in $For_{\mathbf{L}}$, and a chain Δ_n, $n \in \mathbb{N}$, of theories built as follows:

$$\Delta_0 = \Gamma$$
$$\Delta_{n+1} = \begin{cases} \Delta_n \cup \{\varphi_n\}, & \text{if } \Delta_n, \varphi_n \nvdash_{\mathbf{L}} \alpha \\ \Delta_n, & \text{otherwise} \end{cases}$$

Let $\Delta = \bigcup_{n\in\mathbb{N}} \Delta_n$. We will show that Δ is relatively maximal with respect

to α in $\mathbf{L}$. First of all, notice that, by an easy induction over the above chain, we can conclude that $\Delta_n \nvdash_{\mathbf{L}} \alpha$, for every $n \in \mathbb{N}$. Moreover, $\Delta \nvdash_{\mathbf{L}} \alpha$. Indeed, if that was not the case, by compactness there would be some finite $\Delta^{\mathsf{fin}} \subseteq \Delta$ such that $\Delta^{\mathsf{fin}} \vdash_{\mathbf{L}} \alpha$. But then, using cut, there would be some $\Delta_m \supseteq \Delta^{\mathsf{fin}}$ such that $\Delta_m \vdash_{\mathbf{L}} \alpha$, and that is impossible. Now, consider some $\beta \notin \Delta$. That β must be such that $\beta = \varphi_n$, for some n. Thus $\beta \notin \Delta_{n+1}$, given reflexivity and $\Delta_{n+1} \subseteq \Delta$. So, $\Delta_{n+1} = \Delta_n$ and $\Delta_n, \beta \vdash_{\mathbf{L}} \alpha$, by construction. Once $\Delta_n \subseteq \Delta$, we are bound to conclude by monotonicity that $\Delta, \beta \vdash_{\mathbf{L}} \alpha$. ■

We can also prove that:

LEMMA 58. Any relatively maximal set of formulas is a closed theory.

Proof. Given a set of formulas Δ that is relatively maximal with respect to a formula α, we have to check that $\Delta \vdash_{\mathbf{mbC}} \beta$ iff $\beta \in \Delta$. From right to left is obvious by reflexivity. From left to right, given some $\beta \notin \Delta$ we have that (a) $\Delta \nvdash_{\mathbf{mbC}} \alpha$ and (b) $\Delta, \beta \vdash_{\mathbf{mbC}} \alpha$, since Δ is relatively maximal with respect to α. But then, from (a) and (b) we conclude, using cut, that $\Delta \nvdash_{\mathbf{mbC}} \beta$. ■

LEMMA 59. Let $\Delta \cup \{\alpha\}$ be a set of formulas in For° such that Δ is relatively maximal with respect to α in **mbC**. Then:

(i) $(\beta \wedge \gamma) \in \Delta$ iff $\beta \in \Delta$ and $\gamma \in \Delta$.

(ii) $(\beta \vee \gamma) \in \Delta$ iff $\beta \in \Delta$ or $\gamma \in \Delta$.

(iii) $(\beta \rightarrow \gamma) \in \Delta$ iff $\beta \notin \Delta$ or $\gamma \in \Delta$.

(iv) $\beta \notin \Delta$ implies $\neg\beta \in \Delta$.

(v) $\circ\beta \in \Delta$ implies $\beta \notin \Delta$ or $\neg\beta \notin \Delta$.

Proof. The closure guaranteed by Lemma 58 will be used to prove each of the above items.
Item (i) is proven from closure, axioms (Ax3), (Ax4), (Ax5) and (MP).
Item (ii) follows from closure, axioms (Ax6), (Ax7), (Ax8) and (MP).
Item (iii) from closure, (ii), axioms (Ax1), (Ax9) and (MP).
Item (iv) from closure, axiom (Ax10) and (MP).
For item (v), suppose $\beta \in \Delta$ and $\neg\beta \in \Delta$. Then, from closure, (bc1) and relative maximality, we conclude that $\circ\beta \notin \Delta$. ■

COROLLARY 60. The characteristic function of a relatively maximal set of formulas in **mbC** defines an **mbC**-valuation.

Proof. Let Δ be a set of formulas relatively maximal with respect to α and define a function v: $For^{\circ} \longrightarrow \mathbf{2}$ such that, for any formula β in For°, $v(\beta) = 1$ iff $\beta \in \Delta$. Using the previous lemma it is easy to see that v satisfies clauses $(v1)$ to $(v5)$ of Definition 54. ■

THEOREM 61. [Completeness] Let $\Gamma \cup \{\alpha\}$ be a set of formulas in For°. Then: $\Gamma \models_{\mathbf{mbC}} \alpha$ implies $\Gamma \vdash_{\mathbf{mbC}} \alpha$.

Proof. Given a formula α in For° such that $\Gamma \not\vdash_{\mathbf{mbC}} \alpha$ one may, by the Lindenbaum-Asser argument, extend Γ to a set Δ that is relatively maximal with respect to α. As $\Delta \not\vdash_{\mathbf{mbC}} \alpha$, then $\alpha \notin \Delta$, because of (Con1). By Corollary 60, the characteristic function v of Δ is an **mbC**-valuation such that, for any $\beta \in \Delta$, $v(\beta) = 1$, while $v(\alpha) = 0$. So, $\Delta \not\models_{\mathbf{mbC}} \alpha$, and in particular $\Gamma \not\models_{\mathbf{mbC}} \alpha$. ■

Using the bivaluation semantics for **mbC**, we obtain easy semantical proofs of several remarkable features of **mbC** (see Theorem 64 below). Previous to do this, we need to show how it is possible to construct an **mbC**-valuation satisfying a given set of requirements.

DEFINITION 62. Let the mapping ℓ: $For^\circ \longrightarrow \mathbb{N}$ denote the *complexity measure* defined over the signature Σ°, by: $\ell(p) = 0$, for $p \in \mathcal{P}$; $\ell(\varphi \# \psi) = \ell(\varphi) + \ell(\psi) + 1$, for $\# \in \{\wedge, \vee, \rightarrow\}$; $\ell(\neg\varphi) = \ell(\varphi) + 1$; and $\ell(\circ\varphi) = \ell(\varphi) + 2$. ■

LEMMA 63. Let v_0: $\mathcal{P} \cup \{\neg p : p \in \mathcal{P}\} \longrightarrow \mathbf{2}$ be a mapping such that $v_0(\neg p) = 1$ whenever $v_0(p) = 0$ (for $p \in \mathcal{P}$). Then, there exists an **mbC**-valuation v: $For^\circ \longrightarrow \mathbf{2}$ extending v_0, that is, such that $v(\varphi) = v_0(\varphi)$ for every $\varphi \in \mathcal{P} \cup \{\neg p : p \in \mathcal{P}\}$.

Proof. We will define the value of $v(\varphi)$ while doing an induction on the complexity $\ell(\varphi)$ of a formula $\varphi \in For^\circ$. Thus, we begin by setting $v(\varphi) = v_0(\varphi)$ for every $\varphi \in \mathcal{P} \cup \{\neg p : p \in \mathcal{P}\}$, and $v(p\#q)$ is defined according to clauses $(v1)$–$(v3)$ of Definition 54, for $\# \in \{\wedge, \vee, \rightarrow\}$ and $p, q \in \mathcal{P}$. This completes the definition of $v(\varphi)$ for every $\varphi \in For^\circ$ such that $\ell(\varphi) \leq 1$. Suppose now that $v(\varphi)$ has been defined for every $\varphi \in For^\circ$ such that $\ell(\varphi) \leq n$ (for $n \geq 1$) and let $\varphi \in For^\circ$ such that $\ell(\varphi) = n + 1$. If $\varphi = (\psi_1 \# \psi_2)$ for $\# \in \{\wedge, \vee, \rightarrow\}$ then $v(\varphi)$ is defined according to $(v1)$–$(v3)$. If $\varphi = \neg\psi$ then we define $v(\varphi) = 1$, if $v(\psi) = 0$, and $v(\varphi)$ is defined arbitrarily, otherwise. Finally, if $\varphi = \circ\psi$ then $v(\varphi) = 0$, if $v(\psi) = v(\neg\psi) = 1$, and $v(\varphi)$ is defined arbitrarily otherwise. It is clear that v is an **mbC**-valuation that extends the mapping v_0. ■

THEOREM 64. The connectives $\wedge, \vee$ and $\rightarrow$ are not interdefinable as in the classical case. Indeed, the following rule holds good in **mbC**:

(i) $(\neg\alpha \rightarrow \beta) \Vdash (\alpha \vee \beta)$,

but none of the following rules hold in **mbC**:

(ii) $(\alpha \vee \beta) \Vdash (\neg\alpha \rightarrow \beta)$;
(iii) $\neg(\neg\alpha \rightarrow \beta) \Vdash \neg(\alpha \vee \beta)$;
(iv) $\neg(\alpha \vee \beta) \Vdash \neg(\neg\alpha \rightarrow \beta)$;

(v) $(\alpha \rightarrow \beta) \Vdash \neg(\alpha \wedge \neg\beta)$;
(vi) $\neg(\alpha \wedge \neg\beta) \Vdash (\alpha \rightarrow \beta)$;
(vii) $\neg(\alpha \rightarrow \beta) \Vdash (\alpha \wedge \neg\beta)$;
(viii) $(\alpha \wedge \neg\beta) \Vdash \neg(\alpha \rightarrow \beta)$;
(ix) $\neg(\neg\alpha \wedge \neg\beta) \Vdash (\alpha \vee \beta)$;
(x) $(\alpha \vee \beta) \Vdash \neg(\neg\alpha \wedge \neg\beta)$;
(xi) $\neg(\neg\alpha \vee \neg\beta) \Vdash (\alpha \wedge \beta)$;
(xii) $(\alpha \wedge \beta) \Vdash \neg(\neg\alpha \vee \neg\beta)$.

Proof. (i) Let v be an **mbC**-valuation such that $v(\alpha \vee \beta) = 0$. Then $v(\alpha) = 0 = v(\beta)$ and so $v(\neg\alpha) = 1$. Therefore $v(\neg\alpha) = 1$ and $v(\beta) = 0$, that is, $v(\neg\alpha \rightarrow \beta) = 0$. This shows that $(\neg\alpha \rightarrow \beta) \vDash_{\mathbf{mbC}} (\alpha \vee \beta)$. The result for $\vdash_{\mathbf{mbC}}$ follows from Theorem 61.
(ii) Consider a mapping $v_0 \colon \mathcal{P} \cup \{\neg p : p \in \mathcal{P}\} \longrightarrow \mathbf{2}$ such that $v_0(p_0) = 1 = v_0(\neg p_0)$, $v_0(p_1) = 0$ and $v_0(\varphi)$ is defined arbitrarily otherwise. Let v be an **mbC**-valuation extending v_0 (check the Lemma 63). Then $v(p_0 \vee p_1) = 1$ but $v(\neg p_0 \rightarrow p_1) = 0$. This shows that $(p_0 \vee p_1) \nvDash_{\mathbf{mbC}} (\neg p_0 \rightarrow p_1)$. The result for $\vdash_{\mathbf{mbC}}$ follows from Theorem 56.
The remainder of the proof is analogous. ■

EXAMPLE 65. The first **LFI** ever to receive an interpretation in terms of bivaluation semantics was the logic C_1 of Example 28 (cf. [da Costa and Alves, 1977]). The original set of clauses characterizing the C_1-valuations is the following:

$(vC1)$ $v(\alpha_1 \wedge \alpha_2) = 1$ iff $v(\alpha_1) = 1$ and $v(\alpha_2) = 1$;
$(vC2)$ $v(\alpha_1 \vee \alpha_2) = 1$ iff $v(\alpha_1) = 1$ or $v(\alpha_2) = 1$;
$(vC3)$ $v(\alpha_1 \rightarrow \alpha_2) = 1$ iff $v(\alpha_1) = 0$ or $v(\alpha_2) = 1$;
$(vC4)$ $v(\neg\alpha) = 0$ implies $v(\alpha) = 1$;
$(vC5)$ $v(\neg\neg\alpha) = 1$ implies $v(\alpha) = 1$;
$(vC6)$ $v(\circ\beta) = v(\alpha \rightarrow \beta) = v(\alpha \rightarrow \neg\beta) = 1$ implies $v(\alpha) = 0$;
$(vC7)$ $v(\circ(\alpha \# \beta)) = 0$ implies $v(\circ\alpha) = 0$ or $v(\circ\beta) = 0$, for $\# \in \{\wedge, \vee, \rightarrow\}$,

where, as usual, $\circ\alpha$ abbreviates the formula $\neg(\alpha \wedge \neg\alpha)$. ■

*3.4 Possible-translations semantics for **LFI**s*

Notwithstanding the fact that the completeness proof by means of bivaluations for **LFI**s is simple to obtain, this semantics does not do a good job in explaining intrinsic singularities of such logics. In particular, it is not obvious right from the definition of the bivaluation semantics for **mbC** (Definition 54) that this logic is *decidable*. A decision procedure can be obtained with some further effort, however, by adapting the well-known

procedure of truth-tables, or 'matrices', into a procedure of 'quasi matrices' (check for instance [da Costa and Alves, 1977] and [da Costa *et al.*, 1995]). At any rate, bivaluation semantics may be very useful as a technical device that helps in simplifying the completeness proof with respect to possible-translations semantics that we present in this subsection, as well as in defining two-signed tableaux for our logics, as it will be illustrated in the next section. Possible-translations semantics were introduced in [Carnielli, 1990]; for a study of their scope and for formal definitions related to them check [Marcos, 2004]. Of course, the notion of *translation* between a logic **L**1 and a logic **L**2 is essential here (recall Definition 31).

Consider now the following 3-valued truth-tables, where T and t are the designated values:

$\wedge$	T	t	F
T	t	t	F
t	t	t	F
F	F	F	F

$\vee$	T	t	F
T	t	t	t
t	t	t	t
F	t	t	F

$\rightarrow$	T	t	F
T	t	t	F
t	t	t	F
F	t	t	t

	$\neg_1$	$\neg_2$	$\circ_1$	$\circ_2$
T	F	F	t	F
t	F	t	F	F
F	T	t	t	F

In order to provide interpretations to the connectives of **mbC** by means of possible-translations semantics one should first understand these truth-tables. The truth-value t may be interpreted as 'true by default', or 'true by lack of evidence to the contrary', and T and F are, as usual, 'true' and 'false'. The truth-tables for conjunction, disjunction and implication never return the value T, so, in principle, one is never absolutely sure about the truth-status of some compound sentences. There are two distinct interpretations for negation, $\neg$, and for the consistency operator, $\circ$. The basic intuition is the idea of *multiple scenarios* concerning the dynamics of evaluation of propositions: One may think that there are two kinds of situations concerning non-true propositions with respect to successive moments of time. In the first situation, a true-by-default proposition is treated as a true proposition with respect to the negation $\neg_1$. In the other situation, one can consider the case in which the negation of any other value than 'true' becomes true-by-default — this is expressed by the negation $\neg_2$. On what concerns the consistency operator $\circ$, the first interpretation $\circ_1$ only considers as true-by-default the 'classical' values T and F, while $\circ_2$ assigns falsehood to every truth-value.

The above collection of truth-tables, which we call $\mathcal{M}_0$, will be used to provide the desired semantics for **mbC**. Now, considering the algebra

$For_{\mathcal{M}_0}$ of formulas generated by $\mathcal{P}$ over the signature of $\mathcal{M}_0$, let's define the set TR_0 of all mappings $*$: $For^\circ \longrightarrow For_{\mathcal{M}_0}$ subjected to the following restrictive clauses:

$(tr0)$ $p^* = p$, if $p \in \mathcal{P}$;

$(tr1)$ $(\alpha \# \beta)^* = (\alpha^* \# \beta^*)$, for all $\# \in \{\wedge, \vee, \rightarrow\}$;

$(tr2)$ $(\neg\alpha)^* \in \{\neg_1\alpha^*, \neg_2\alpha^*\}$;

$(tr3)$ $(\circ\alpha)^* \in \{\circ_1\alpha^*, \circ_2\alpha^*, \circ_1(\neg\alpha)^*\}$.

We say the pair $\mathrm{PT}_0 = \langle \mathcal{M}_0, \mathrm{TR}_0 \rangle$ is a *possible-translations semantical structure for* **mbC**. If $\vDash_{\mathcal{M}_0}$ denotes the consequence relation in $\mathcal{M}_0$, and $\Gamma \cup \{\alpha\}$ is a set of formulas of **mbC**, the associated PT-consequence relation, $\models_{\mathrm{PT}_0}$, is defined by setting:

$$\Gamma \models_{\mathrm{PT}_0} \alpha \text{ iff } \Gamma^* \vDash_{\mathcal{M}_0} \alpha^* \text{ for all translations } * \text{ in } \mathrm{TR}_0.$$

We will call *possible translation* of a formula α any image of it through some mapping in TR_0. One can immediately check the following:

THEOREM 66. [Soundness] Let $\Gamma \cup \{\alpha\}$ be a set of formulas of **mbC**. Then $\Gamma \vdash_{\mathbf{mbC}} \alpha$ implies $\Gamma \models_{\mathrm{PT}_0} \alpha$.

Proof. It is sufficient to check that the (finite) collection of all possible translations of each axiom produces tautologies in the truth-tables of $\mathcal{M}_0$ and that all possible translations of the rule (MP) preserve validity. The verification is immediate, and we leave it as exercise to the reader. ■

As a corollary of the above result, we see that each mapping in TR_0 defines in fact a translation (recall Definition 31) from **mbC** to the logic defined by $\mathcal{M}_0$.

In order to prove completeness, now, our strategy will be to show that each **mbC**-valuation v determines a translation $*$ and a 3-valued valuation w defined in the usual way over the truth-tables of $\mathcal{M}_0$ such that, for every formula α of **mbC**,

$$w(\alpha^*) \in \{T, t\} \text{ iff } v(\alpha) = 1$$

and thus rely on the completeness proof for the bivaluation semantics of **mbC**.

Recall the definition of complexity $\ell(\alpha)$ of a formula $\alpha \in For^\circ$ introduced in Definition 62. The following result comes from [Marcos, 2005f]:

THEOREM 67. [Representability] Given an **mbC**-valuation v there is a translation $*$ in TR_0 and a valuation w in $\mathcal{M}_0$ such that, for every formula α in **mbC**:

$$w(\alpha^*) = T \text{ implies } v(\neg\alpha) = 0; \text{ and}$$
$$w(\alpha^*) = F \text{ iff } v(\alpha) = 0.$$

Proof. For $p \in \mathcal{P}$ define the valuation w as follows:

$$\begin{array}{lll} w(p) = F & \text{if} & v(p) = 0; \\ w(p) = T & \text{if} & v(p) = 1 \text{ and } v(\neg p) = 0; \\ w(p) = t & \text{if} & v(p) = 1 \text{ and } v(\neg p) = 1. \end{array}$$

Such w can be homomorphically extended to the algebra $For_{\mathcal{M}_0}$. We define the translation mapping $*$ as follows:

1. $p^* = p$, if $p \in \mathcal{P}$;
2. $(\alpha \# \beta)^* = (\alpha^* \# \beta^*)$, for $\# \in \{\wedge, \vee, \rightarrow\}$;
3. $(\neg\alpha)^* = \neg_1\alpha^*$, if $v(\neg\alpha) = 0$ or $v(\alpha) = v(\neg\neg\alpha) = 0$;
4. $(\neg\alpha)^* = \neg_2\alpha^*$, otherwise;
5. $(\circ\alpha)^* = \circ_2\alpha^*$, if $v(\circ\alpha) = 0$;
6. $(\circ\alpha)^* = \circ_1(\neg\alpha)^*$, if $v(\circ\alpha) = 1$ and $v(\neg\alpha) = 0$;
7. $(\circ\alpha)^* = \circ_1\alpha^*$, otherwise.

Note that the mapping $*$ is well-defined, given the definition of **mbC** (see Definition 54). The proof is now done by induction on the complexity measure $\ell(\alpha)$ of a formula α. Details are left to the reader. ■

THEOREM 68. [Completeness] Let $\Gamma \cup \{\alpha\}$ be a set of formulas in **mbC**. Then $\Gamma \models_{\text{PT}_0} \alpha$ implies $\Gamma \vdash_{\mathbf{mbC}} \alpha$.

Proof. Suppose that $\Gamma \models_{\text{PT}_0} \alpha$, and suppose that v is an **mbC**-valuation such that $v(\Gamma) \subseteq \{1\}$. By Theorem 67, there is a translation $*$ and a 3-valued valuation w such that, for every formula β, $w(\beta^*) \in \{T, t\}$ iff $v(\beta) = 1$. From this, $w(\Gamma^*) \subseteq \{T, t\}$ and so $w(\alpha^*) \in \{T, t\}$, because $\Gamma \models_{\text{PT}_0} \alpha$. Then $v(\alpha) = 1$. To wit: For every **mbC**-valuation v, $v(\Gamma) \subseteq \{1\}$ implies $v(\alpha) = 1$. Using the completeness of **mbC** with respect to **mbC**-valuations we obtain that $\Gamma \vdash_{\mathbf{mbC}} \alpha$ as desired. ■

It is now easy to check validity for inferences in **mbC**, as shown in the following example.

EXAMPLE 69. We will prove that $\circ\alpha \vdash_{\mathbf{mbC}} \neg(\neg\alpha \wedge \alpha)$ using possible-translations semantics. We have that, for any translation $*$ in T_{R_0},

$$(\circ\alpha)^* \in \{\circ_1(\alpha^*), \circ_2(\alpha^*), \circ_1\neg_1(\alpha^*), \circ_1\neg_2(\alpha^*)\},$$

$$(\neg(\neg\alpha \wedge \alpha))^* \in \{\neg_i(\neg_j(\alpha^*) \wedge \alpha^*) : i, j \in \{1, 2\}\}.$$

Let $*$ be a translation in T_{R_0}, w be a valuation in $\mathcal{M}_0$, and $D = \{T, t\}$. Let $x = w(\alpha^*)$, $y = w((\circ\alpha)^*)$ and $z = w((\neg(\neg\alpha \wedge \alpha))^*)$, and suppose that $y \in D$; this rules out the translation $(\circ\alpha)^* = \circ_2(\alpha^*)$ because $\circ_2(x) \notin D$. In order to prove that $z \in D$ we have the following cases:

1. $(\circ\alpha)^* = \circ_1(\alpha^*)$. Then $\circ_1(x) \in D$, thus $x \in \{T, F\}$.

(a) $x = T$. Then $\neg_j(x) = F$ ($j \in \{1,2\}$) and so $\neg_i(\neg_j(x) \wedge x) \in D$ for $i, j \in \{1,2\}$.

(b) $x = F$. Then $(\neg_j(x) \wedge x) = F$ ($j \in \{1,2\}$) and so $\neg_i(\neg_j(x) \wedge x) \in D$ for $i, j \in \{1,2\}$.

2. $(\circ\alpha)^* = \circ_1\neg_1(\alpha^*)$. Then $\circ_1\neg_1(x) \in D$, thus $\neg_1(x) \in \{T, F\}$ and $z = \neg_i(\neg_1(x) \wedge x)$.
   (a) $\neg_1(x) = T$. Then $x = F$ and the proof is as in (1b).
   (b) $\neg_1(x) = F$. In this case the proof is as in (1a).
3. $(\circ\alpha)^* = \circ_1\neg_2(\alpha^*)$. Then, given $\circ_1\neg_2(x) \in D$, we have $\neg_2(x) \in \{T, F\}$ and $z = \neg_i(\neg_2(x) \wedge x)$. From the truth-table for $\neg_2$ we obtain that $\neg_2(x) = F$, and the proof is as in (1a).

This proves the desired result. On the other hand, we may prove that the converse $\neg(\neg\alpha \wedge \alpha) \vdash_{\mathbf{mbC}} \circ\alpha$ is not true in **mbC**, as announced in Theorem 49. Using the same notation as above for a given translation $*$ in TR_0 and a valuation w in $\mathcal{M}_0$, it is enough to consider α as a propositional variable p, and choose $*$ and w such that $x = F$, and $(\circ\alpha)^* = \circ_2(\alpha^*)$. Then $z \in D$ and $y = F$. For yet some other counter-models to that inference, take $x = t$, $(\neg(\neg\alpha \wedge \alpha))^* = \neg_2(\neg_2(\alpha^*) \wedge \alpha^*)$ and $(\circ\alpha)^* \in \{\circ_1(\alpha)^*, \circ_1(\neg_2\alpha)^*\}$. ■

Possible-translations semantics offer an immediate decision procedure for any logic **L** that is complete with respect to a possible-translations semantical structure $\mathrm{PT} = \langle \mathcal{M}, \mathrm{TR} \rangle$ where $\mathcal{M}$ is decidable (and this is the case here, where $\mathcal{M}$ is a finite-valued logic) and TR is recursive. Indeed, given a formula α, if we wish to decide whether it is a theorem of **L**, it is sufficient to consider the (in this case finitely many) possible translations of α, and to check each translated formula using the corresponding semantics of the target logics (in the present case, defined by sets of 3-valued truth-tables). Questions on the complexity of such decision procedures could be readily answered by taking into account the complexity of translations and of the semantics of the target logics. This is a problem of independent interest, since it is immediate to see that the decision procedure of **mbC** is NP-complete, as one might expect: Indeed, there exists a polynomial-time conservative translation from **CPL** into **mbC**, as illustrated in Theorem 74 below.

One can also use possible-translations semantics to help proving important properties about the logics in question.

REMARK 70. Recall from Remark 43 the two explosive negations represented by $\wr\alpha \stackrel{\text{def}}{=} (\neg\alpha \wedge \circ\alpha)$ and $\sim\alpha \stackrel{\text{def}}{=} \alpha \rightarrow (p_0 \wedge (\neg p_0 \wedge \circ p_0))$. Recall again, also, the notion of a classical negation from Definition 8. Now, while it is easy to check that $\sim$ defines a classical negation in **mbC** (the reader can, as an exercise, check that both $(\alpha \vee \sim\alpha)$ and $(\alpha \rightarrow (\sim\alpha \rightarrow \beta))$ are provable / validated by **mbC**), it is also straightforward to check that $\wr$ is not a complementing negation. Indeed, to see that α and $\wr\alpha$ can be simultaneously

false, take some bottom particle $\bot = p \wedge \wr p$ and notice that $w(\bot^*) = F$, for any valuation w in $\mathcal{M}_0$ and any translation $*$ in TR_0. Consider now some translation such that $(\circ p) = \circ_2 p$. In that case, $w((\wr \bot)^*) = F$, for any w. Then, while $\bot \models_{\mathrm{PT}_0} \wr \bot$ certainly holds good, it is not the case that $\models_{\mathrm{PT}_0} \wr \bot$. Notice moreover that, while $\wr \alpha \models_{\mathrm{PT}_0} \sim\alpha$, we have that $\sim\alpha \not\models_{\mathrm{PT}_0} \wr \alpha$. ■

We trust the above features to confirm the importance of possible-translations semantics as a philosophically apt and computationally useful semantical tool for treating not only Logics of Formal Inconsistency but also many other logics in the literature. An remarkable particular case of possible-translations semantics is the so-called *non-deterministic semantics* (cf. [Avron and Lev, 2005]), proposed as an immediate generalization of the notion of a truth-functional semantics (for comparisons between possible-translations semantics and non-deterministic semantics see [Carnielli and Coniglio, 2005]). A 3-valued non-deterministic semantics for the logic **mbC** may be found in [Avron, 2005a] (where this logic is called **B**).

*3.5 Tableau proof systems for **LFI**s*

In this section we will use a very general method to obtain adequate tableau systems for **mbC** and for C_1. The method introduced in [Caleiro *et al.*, 2005b] (check also [Caleiro *et al.*, 2005a]) permits one to obtain an adequate tableau system for any propositional logic which has an adequate semantics given through the so-called 'dyadic valuations'. Such bivaluations have, as usual, values in $\mathbf{2} = \{0, 1\}$ (or, equivalently, in $\{T, F\}$), and are axiomatized by first-order clauses of a certain specific form, explained below.

Briefly, suppose that there is a set of clauses governing a class of bivaluation mappings v: $For \longrightarrow \mathbf{2}$ of the form

$$(v(\varphi_1) = Q_1, \ldots, v(\varphi_n) = Q_n) \Rrightarrow (S_1 | \cdots | S_k)$$

where $n \geq 0$ and $k \geq 0$ and, for every $1 \leq i \leq k$,

$$S_i = (v(\varphi_1^i) = Q_1^i, \ldots, v(\varphi_{r_i}^i) = Q_{r_i}^i),$$

with $Q_i, Q_j^i \in \{T, F\}$ $(1 \leq j \leq r_i)$ and $r_i \geq 1$. If $n = 0$ then $(v(\varphi_1) = Q_1, \ldots, v(\varphi_n) = Q_n)$ is just $\top$; if $k = 0$ then $(S_1 | \cdots | S_k)$ is $\bot$. Commas ',' and bars '|' denote conjunctions and disjunctions, respectively, and '$\Rrightarrow$' denotes implication. Examples of axioms for bivaluations that may be put in this format are provided by the clauses that characterize **mbC**-valuations (cf. Definition 54) and also by those provided by the characteristic bivaluation semantics of da Costa's C_1 (cf. Example 65). For instance, clause $(v5)$ of Definition 54 clearly has the required form:

$(v5)$ $v(\circ\alpha) = T \Rrightarrow (v(\alpha) = F \mid v(\neg\alpha) = F)$

whereas clause $(v3)$ may be split into three clauses of the required form:

(v3.1) $v(\alpha \to \beta) = T \Rrightarrow (v(\alpha) = F \mid v(\beta) = T)$;
(v3.2) $v(\alpha) = F \Rrightarrow v(\alpha \to \beta) = T$;
(v3.3) $v(\beta) = T \Rrightarrow v(\alpha \to \beta) = T$.

It will be convenient in what follows to keep the more complex formulas on the left-hand side of the implication; we thus substitute (v3.2) and (v3.3) by:

(v3.4) $v(\alpha \to \beta) = F \Rrightarrow (v(\alpha) = T, v(\beta) = F)$

The next step in the algorithm described in [Caleiro *et al.*, 2005b] is to 'translate' every clause of the dyadic semantics into a tableau rule by interpreting an equation '$v(\varphi) = Q$' as a signed formula $Q(\varphi)$ (recalling that $Q \in \{T, F\}$). Thus, a clause as above is transformed in a (two-signed) tableau rule of the form:

$$\begin{array}{ccc} & Q_1(\varphi_1), \ldots, Q_n(\varphi_n) & \\ & \diagup \quad \cdots \quad \diagdown & \\ Q_1^1(\varphi_1^1) & & Q_1^k(\varphi_1^k) \\ \vdots & & \vdots \\ Q_{r_1}^1(\varphi_{r_1}^1) & & Q_{r_k}^k(\varphi_{r_k}^k) \end{array}$$

By transforming each clause of the dyadic semantic valuation into a tableau rule, we obtain a tableau system for the given logic. In order to ensure completeness of the tableau system, it is necessary to consider two extra axioms for the bivaluation semantics:

(DV1) $(v(\varphi) = T,\ v(\varphi) = F) \Rrightarrow \bot$;
(DV2) $\top \Rrightarrow (v(\varphi) = T \mid v(\varphi) = F)$.

Axioms (DV1) and (DV2) guarantee that the mapping respecting them is a bivaluation v: $For \longrightarrow \mathbf{2}$. The translation of axiom (DV1) gives us the usual closure condition for a branch in a given tableau. On the other hand, (DV2) gives us the following branching tableau rule, R_b:

$$\frac{}{T(\varphi) \mid F(\varphi)}$$

As a consequence, the resulting tableau system loses the 'analytic' character. Fortunately, in many important cases this branching rule can be eliminated or at least it can have its scope of application restricted to formulas of a certain format.

We apply next the above technique to obtain an adequate tableau system for the logic **mbC**, based on the bivaluation semantics presented in Definition 54.

EXAMPLE 71. We define an adequate tableau system for **mbC** as follows:

$$\frac{F(\neg X)}{T(X)} \qquad \frac{T(\circ X)}{F(X) \mid F(\neg X)} \qquad \frac{}{T(X) \mid F(X)}$$

$$\frac{T(X_1 \wedge X_2)}{T(X_1),\ T(X_2)} \qquad \frac{F(X_1 \wedge X_2)}{F(X_1) \mid F(X_2)}$$

$$\frac{T(X_1 \vee X_2)}{T(X_1) \mid T(X_2)} \qquad \frac{F(X_1 \vee X_2)}{F(X_1),\ F(X_2)}$$

$$\frac{T(X_1 \rightarrow X_2)}{F(X_1) \mid T(X_2)} \qquad \frac{F(X_1 \rightarrow X_2)}{T(X_1),\ F(X_2)}$$

■

Observe that, except for the branching rule R_b, all other rules are analytic in the sense that the consequences are always less complex than the premises (recall that, as in Definition 62, $\ell(\circ\alpha) = \ell(\alpha) + 2$ and $\ell(\neg\alpha) = \ell(\alpha) + 1$), and they contain in each case only subformulas of the premise. The results proven in [Caleiro *et al.*, 2005b] guarantee that the tableau system defined above is sound and complete for **mbC**.

Another nice application of the techniques described above is the definition of a tableau system for the historical **dC**-system C_1 (see Definition 28).

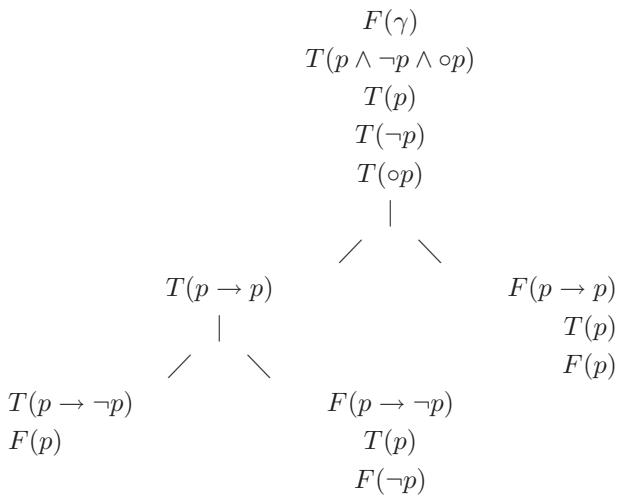
EXAMPLE 72. Recall from Example 65 the characteristic bivaluation semantics for the logic C_1. Those clauses of course may be formally rewritten as axioms of a dyadic semantics, using '|', '$\Rrightarrow$' and ','. Using the above described method it is immediate to define a complete tableau system associated to these axioms. Consider indeed all the rules of the tableau system for **mbC** in Example 71 — except for the rule concerning $\circ$, since it does not correspond to any axiom of a C_1-valuation — and add the following further rules:

$$\frac{T(\neg\neg X)}{T(X)} \qquad \frac{F(\circ(X_1 \# X_2))}{F(\circ X_1) \mid F(\circ X_2)} \qquad \frac{T(\circ X_2),\ T(X_1 \rightarrow X_2),\ T(X_1 \rightarrow \neg X_2)}{F(X_1)}$$

where $\# \in \{\wedge, \vee, \rightarrow\}$ and $\circ X$ abbreviates the formula $\neg(X \wedge \neg X)$. Comparing this tableau system with the one defined in [Carnielli and Lima-Marques, 1992], we notice that the present system does not allow for loops. Although the looping rules proposed in the latter paper often permit one to obtain somewhat conciser tableau proofs, what we have here is a generic method that *automatically generates* a complete set of tableau rules (though not necessarily the most convenient one). ■

It is worth reinforcing that the branching rule R_b is essential, above, in order to obtain completeness. This rule is not strictly analytic, but it can be bounded in certain cases so as to guarantee the termination of the decidable tableau procedure. In particular, the variables occurring in the formula X must be contained in the finite collection of variables in the tableau branch.

EXAMPLE 73. Consider the tableau system for C_1 given in Example 72 and let γ be the formula $\neg(p \wedge (\neg p \wedge \circ p))$, where p is a propositional variable. The formula γ is a thesis of C_1. However, it is easy to see that no C_1-tableau for the set $\{F(\gamma)\}$ can close without using the rule R_b. We show below a closed tableau for $\{F(\gamma)\}$ that uses R_b twice.



■

This example suggests that, in general, it is not possible to eliminate R_b if one wishes to obtain completeness. This holds even in case the tableau system satisfies the subformula property, as in Example 73. In certain cases R_b can be eliminated if we have, for instance, looping rules as in [Carnielli and Lima-Marques, 1992]. For the case of C_1 the tableau system treated in the latter paper uses the looping rule:

$$\frac{T(\neg X)}{F(X) \mid F(\circ X)},$$

while our present formulation has no rule for analyzing $T(\neg X)$.

3.6 Talking about classical logic

When attempting to compare the inferential power of two logics, one often finds difficulties because those logics might not be 'talking about the same thing'. For instance, **mbC** is written in a richer signature than that of

CPL, and so these two logics might seem hard to compare. However, as we have seen in Remark 30, it is possible to linguistically extend **CPL** by the addition of a consistency-like connective. The 'classical truth-tables' for this connective, however, will be such that $\circ(x) = 1$ for every x. Clearly, despite being gently explosive, the resulting logic **eCPL** does not define an **LFI**, given that it is not paraconsistent. It is, indeed, a *consistent* logic (recall Definition 4). Now, **mbC** may be characterized as a deductive fragment of **eCPL**, because all axioms of **mbC** are validated by the truth-tables of **eCPL**. Since **mbC** is a fragment thus of (an alternative formulation of) classical logic, we can conclude that **mbC** is a non-contradictory and non-trivial logic. On the other hand, however, we will show in this subsection that there are several ways of encoding each inference of **CPL** within **mbC**.

First of all, recall the **DAT**s from Remark 26, the Derivability Adjustment Theorems that described how the **LFI**s could be used to recover consistent reasoning by the addition in each case of a convenient number of consistency assumptions. In particular, in logics such as **mbC**, **C**-systems based on classical logic, it should be clear how classical reasoning may be recovered. For each classical rule that is lost by paraconsistency, such as *reductio* and contraposition in items (ii) and (iii) of Theorem 38, there is an adjusted version of the same rule that is gained, as illustrated in Theorem 48. Indeed, it is now easy to give a semantical proof that:

$$\forall\Gamma\forall\gamma\exists\Delta(\Gamma \Vdash_{\mathbf{eCPL}} \gamma \text{ iff } \circ(\Delta), \Gamma \Vdash_{\mathbf{mbC}} \gamma).$$

Now, besides the **DAT**s, there might well be other more direct ways of recovering consistent reasoning from inside a given **LFI**. We will in the following show how this can be done through the use of a conservative translation (recall Definition 31), without the addition of further assumptions to the set of premises of a given inference.

Except for negation and for the consistency connective, all other connectives of **mbC** have a classic-like behavior. The key for the next result will be, thus, to make use of the classical negation $\sim$ that can be defined within **mbC** (cf. Remark 70) by setting $\sim\alpha \stackrel{\text{def}}{=} \alpha \rightarrow (p_0 \wedge (\neg p_0 \wedge \circ p_0))$, in order to recover all classical inferences.

THEOREM 74. Let For° be the algebra of formulas for the signature Σ° of **mbC**. There is a mapping t_1: $For \longrightarrow For^\circ$ that conservatively translates **CPL** inside of **mbC**, that is, for every $\Gamma \cup \{\alpha\} \subseteq For$:

$$\Gamma \vdash_{\mathbf{CPL}} \alpha \text{ iff } t_1(\Gamma) \vdash_{\mathbf{mbC}} t_1(\alpha).$$

Proof. Define the mapping t_1 recursively as follows:

1. $t_1(p) = p$, for every $p \in \mathcal{P}$;
2. $t_1(\gamma \# \delta) = t_1(\gamma) \# t_1(\delta)$, if $\# \in \{\wedge, \vee, \rightarrow\}$;
3. $t_1(\neg\gamma) = \sim t_1(\gamma)$.

Since both **CPL** and **mbC** are compact and have a deductive implication, and considering that t_1 preserves implications, it suffices to prove that:

$$\vdash_{\mathbf{CPL}} \alpha \text{ iff } \vdash_{\mathbf{mbC}} t_1(\alpha)$$

for every $\alpha \in For$.

That $\vdash_{\mathbf{CPL}} \alpha$ implies $\vdash_{\mathbf{mbC}} t_1(\alpha)$ is an easy consequence of the fact that $\sim$ is a classical negation within **mbC** and from the definition of the translation mapping t_1. Let's now check that $\vdash_{\mathbf{mbC}} t_1(\alpha)$ implies $\vdash_{\mathbf{CPL}} \alpha$. Consider the classical truth-tables for the classical connectives, and define $\circ(x) = 1$ for all x. Then $\neg\alpha$ and $\sim\alpha$ take the same value and so $t_1(\alpha)$ and α take the same value in this semantics. Therefore, if $t_1(\alpha)$ is a theorem of **mbC** then $t_1(\alpha)$ is valid for the above truth-tables and so α is valid using classical truth-tables. Thus, α is a theorem of **CPL**, by the completeness of classical logic. ■

In view of the last theorem, and as it was already mentioned, it is clear that **mbC** (originally defined as a *deductive fragment* of **eCPL**) can also be seen as an *extension* of **CPL**, if we employ an appropriate signature which contains two symbols for negation: $\sim$ for the classical one, and provided $\neg$ for the paraconsistent one, provided that the axioms defining $\sim$ in terms of the other connectives are added to the new version of **mbC**.

In what follows, and in stronger logics than **mbC**, we will see yet other ways of recovering classical inferences inside our **LFI**s (check Theorems 96, 112 and 113).

4 A RICHER **LFI**

*4.1 The system **mCi**, and its significance*

In Remark 45 we have mentioned the possibility of defining in **mbC** an inconsistency connective that is dual to its native consistency connective. This could be done by setting $\bullet\alpha \stackrel{\text{def}}{=} \sim\circ\alpha$, where $\sim\alpha \stackrel{\text{def}}{=} \alpha \rightarrow (\beta \wedge (\neg\beta \wedge \circ\beta))$ (for an arbitrary β) is a classical negation. Now, how could we enrich **mbC** so as to be able to define the inconsistency connective by using the paraconsistent negation instead of the classical $\sim$, that is, by setting $\bullet\alpha \stackrel{\text{def}}{=} \neg\circ\alpha$? This is exactly what will be done in this subsection by extending **mbC** into the logic **mCi**. In fact, **mCi** will reveal to be a logic that can be presented in terms of either $\circ$ or $\bullet$ as primitive connectives. Moreover, $\bullet\alpha$ and $\neg\circ\alpha$ will be inter-translatable, and the same will happen with $\circ\alpha$ and $\neg\bullet\alpha$, as proven in Theorem 98.

From Theorem 49(i) we know that $\alpha \wedge \neg\alpha \vdash_{\mathbf{mbC}} \neg\circ\alpha$. The converse property (which does not hold in **mbC**) will be the first additional axiom we will add to **mbC** in upgrading this logic. On the other hand, we wish

that formulas of the form $\neg\circ\alpha$ 'behave classically', and we wish to obtain in fact a logic that is controllably explosive in contact with formulas of the form $\neg^n\circ\alpha$, where $\neg^0\alpha \stackrel{\text{def}}{=} \alpha$ and $\neg^{n+1}\alpha \stackrel{\text{def}}{=} \neg\neg^n\alpha$. Any formula of the form $\neg^n\circ\alpha$ would thus be assumed to 'behave classically', and $\{\neg^n\circ\alpha, \neg^{n+1}\circ\alpha\}$ would be an explosive theory. This desideratum leads us into considering the following (cf. [Marcos, 2005f]):

DEFINITION 75. The logic **mCi** is obtained from **mbC** (recall Definition 42) by the addition of the following axiom schemas:
(ci) $\neg\circ\alpha \rightarrow (\alpha \wedge \neg\alpha)$
(cc)$_n$ $\circ\neg^n\circ\alpha \qquad (n \geq 0)$
To the above axiomatization we add the definition by abbreviation of an inconsistency connective $\bullet$ by setting $\bullet\alpha \stackrel{\text{def}}{=} \neg\circ\alpha$. ■

Notice that $\neg\circ\alpha$ and $(\alpha \wedge \neg\alpha)$ are equivalent in **mCi**. Clearly every set $\{\neg^n\circ\alpha, \neg^{n+1}\circ\alpha\}$ is explosive in **mCi**, in view of (bc1) and (cc)$_n$. This expresses the 'classical behavior' of formulas of the form $\circ\alpha$ (with respect to the paraconsistent negation). In other words, a formula α in general needs the extra assumption $\circ\alpha$ to 'behave classically', but the formula $\circ\alpha$ and its iterated negations will always 'behave classically'. In Theorem 78 below we will see that $\neg\bullet\alpha$ is equivalent to $\circ\alpha$, and in Definition 97, further on, we will introduce a new formulation of **mCi** that introduces $\bullet$ as a primitive connective. Notice in that case how close is the bond that is established here in between inconsistency and contradictoriness by way of the paraconsistent negation.

We can immediately check that the equivalence in **mCi** between $\neg\circ\alpha$ and $(\alpha \wedge \neg\alpha)$ is in fact logically weaker than the identification of $\circ\alpha$ and $\neg(\alpha \wedge \neg\alpha)$ assumed in C_1 (recall also Theorem 49(iii)–(iv)) since the latter formula implies the former, in **mCi**, but the converse is not true.

THEOREM 76. This rule holds good in **mCi**:
(i) $\neg\circ\alpha \vdash_{\mathbf{mCi}} (\alpha \wedge \neg\alpha)$,
but the following rules do not hold:
(ii) $\neg(\alpha \wedge \neg\alpha) \vdash_{\mathbf{mCi}} \circ\alpha$;
(iii) $\neg(\neg\alpha \wedge \alpha) \vdash_{\mathbf{mCi}} \circ\alpha$.

Proof. Item (i) is obvious. In order to prove that (ii) and (iii) do not hold in **mCi**, observe that **mCi** is sound for the truth-tables of **LFI1** (see Examples 17 and 18), where 0 is the only non-designated value. Then it is enough to check that (ii) and (iii) have counter-models in such a truth-functional semantics. ■

It should be clear that, even though in **mCi** there is a formula in the classical language *For* (namely, the formula $(\alpha \wedge \neg\alpha)$) that is equivalent to a formula that expresses inconsistency (the formula $\bullet\alpha$), there is no formula in the classical language that can express consistency in **mCi**. We also have the following:

THEOREM 77. (i) $\neg(\alpha \wedge \neg\alpha)$ and $\neg(\neg\alpha \wedge \alpha)$ are not top particles in **mCi**.
(ii) $\circ\alpha$ and $\neg\circ\alpha$ are not bottom particles.
(iii) The schemas $(\alpha \rightarrow \neg\neg\alpha)$ and $(\neg\neg\alpha \rightarrow \alpha)$ are not provable in **mCi**.

Proof. Items (i), (ii) and the first part of item (iii) can be checked using again the truth-tables of $\mathbf{P}^1$, enriched with the (definable) truth-table for $\circ$ (Example 19), and using the fact that **mCi** is sound for such a semantics. For the second part of item (iii) one could use for instance the bivaluation semantics of **mCi** (see Example 90). ■

It is straightforward to check the following properties of **mCi**:

THEOREM 78. The following rules hold good in **mCi**:
(i) $\neg\neg\circ\alpha \vdash_{\mathbf{mCi}} \circ\alpha$;
(ii) $\circ\alpha \vdash_{\mathbf{mCi}} \neg\neg\circ\alpha$;
(iii) $\circ\alpha, \neg\circ\alpha \vdash_{\mathbf{mCi}} \beta$;
(iv) $(\Gamma, \beta \vdash_{\mathbf{mCi}} \circ\alpha)$ and $(\Delta, \beta \vdash_{\mathbf{mCi}} \neg\circ\alpha)$ implies $(\Gamma, \Delta \vdash_{\mathbf{mCi}} \neg\beta)$.

Proof. For item (i), from $\neg\neg\circ\alpha$ and $\circ\alpha$ we obviously prove $\circ\alpha$ in **mCi**. On the other hand, from $\neg\neg\circ\alpha$ and $\neg\circ\alpha$ we also prove $\circ\alpha$ in **mCi**, because $\circ\neg\circ\alpha$ and (bc1) are axioms of **mCi**. Using proof-by-cases we conclude that $\neg\neg\circ\alpha \vdash_{\mathbf{mCi}} \circ\alpha$. The other items are proven similarly. Notice in particular how items (i) and (ii) together show that $\neg\bullet\alpha \dashv\vdash_{\mathbf{mCi}} \circ\alpha$ holds good. ■

Item (ii) of Theorem 77 and item (iii) of Theorem 78 guarantee that **mCi** is controllably explosive in contact with $\circ p_0$ (recall Definition 9(iii)). In fact, the following relation between consistency and controllable explosion can be checked:

THEOREM 79. Let **L** be a non-trivial extension of **mCi** such that the implication (involving the axioms of **mCi**) is deductive (recall Definition 6). A schema $\sigma(p_0, \ldots, p_n)$ is provably consistent in **L** if, and only if, **L** is controllably explosive in contact with $\sigma(p_0, \ldots, p_n)$.

Proof. If $\vdash_{\mathbf{L}} \circ\sigma(\alpha_0, \ldots, \alpha_n)$ then, by axiom (bc1),

$$\Gamma, \sigma(\alpha_0, \ldots, \alpha_n), \neg\sigma(\alpha_0, \ldots, \alpha_n) \vdash_{\mathbf{L}} \beta$$

for any choice of Γ and β.

Conversely, assume that $\Gamma, \sigma(\alpha_0, \ldots, \alpha_n), \neg\sigma(\alpha_0, \ldots, \alpha_n) \vdash_{\mathbf{L}} \beta$ for any Γ and β. Since, from (ci), we have that $\neg\circ\sigma(\alpha_0, \ldots, \alpha_n) \vdash_{\mathbf{L}} (\sigma(\alpha_0, \ldots, \alpha_n) \wedge \neg\sigma(\alpha_0, \ldots, \alpha_n))$, then it follows that $\neg\circ\sigma(\alpha_0, \ldots, \alpha_n)$ is a bottom particle. As in the proof of Theorem 46(i) (using here the fact that the original implication of **mCi** is still deductive in **L**), we get $\vdash_{\mathbf{L}} \neg\neg\circ\sigma(\alpha_0, \ldots, \alpha_n)$. By Theorem 78(i), we conclude that $\vdash_{\mathbf{L}} \circ\sigma(\alpha_0, \ldots, \alpha_n)$. ■

Complementing the versions of contraposition mentioned in Theorem 48, we have:

THEOREM 80. Here are some restricted forms of contraposition introduced by **mCi**:

(i) $(\alpha \rightarrow \circ\beta) \vdash_{\mathbf{mCi}} (\neg\circ\beta \rightarrow \neg\alpha)$;
(ii) $(\alpha \rightarrow \neg\circ\beta) \vdash_{\mathbf{mCi}} (\circ\beta \rightarrow \neg\alpha)$;
(iii) $(\neg\alpha \rightarrow \circ\beta) \vdash_{\mathbf{mCi}} (\neg\circ\beta \rightarrow \alpha)$;
(iv) $(\neg\alpha \rightarrow \neg\circ\beta) \vdash_{\mathbf{mCi}} (\circ\beta \rightarrow \alpha)$.

Proof. Item (i). By axiom $(\mathrm{cc})_0$, $\circ\circ\beta$ is a theorem of **mCi**. The result now follows from Theorem 48(iii). The other items are proven similarly. ■

On the other hand, properties such as $(\circ\alpha \rightarrow \beta) \vdash_{\mathbf{mCi}} (\neg\beta \rightarrow \neg\circ\alpha)$ do not hold; this can easily be checked after Corollary 93, to be established below. Notice how the above theorem opens yet another way for the internalization of classical inferences, as discussed in Subsection 3.6.

Recall now the replacement property (RP) discussed in Remark 51. We had already proven in Theorem 52 that (RP) cannot hold in certain paraconsistent extensions of **mbC**. On what concerns its possible validity in paraconsistent extensions of **mCi**, we can now prove that:

THEOREM 81.

(i) The replacement property (RP) is not enjoyed by **mCi**.

The replacement property (RP) cannot hold in any paraconsistent extension of **mCi** in which:

(ii) $\neg(\neg\alpha \wedge \neg\beta) \vdash_{\mathbf{mbC}} (\alpha \vee \beta)$ holds; or
(iii) $(\neg\alpha \vee \neg\beta) \vdash_{\mathbf{mbC}} \neg(\alpha \wedge \beta)$ holds.

Proof. (i) Consider again the first set of truth-tables (with the same set of designated values) used in the proof of Theorem 50.

(ii) Consider the supplementing negation $\wr\alpha = (\neg\alpha \wedge \circ\alpha)$ for **mCi** proposed in Remark 43. By Theorem 78 this last formula is equivalent to $(\neg\alpha \wedge \neg\neg\circ\alpha)$. In Theorem 94, this negation will be shown to behave classically inside this logic. But then, $\neg\wr\alpha \vdash \alpha \vee \neg\circ\alpha$, by hypothesis, and so $\neg\wr\alpha \vdash \alpha$, using axiom (ci), proof-by-cases and conjunction elimination. The rest of the proof now follows exactly like in Theorem 52(ii).

Finally, for item (iii), recall that, from (Ax10), $(\neg\alpha \vee \neg\neg\alpha)$ is a theorem of **mCi**. But then, by hypothesis, $\neg(\alpha \wedge \neg\alpha)$ would also be a theorem. From Theorem 49(ii) and replacement it follows that $\neg\neg\circ\alpha$ is provable, and by Theorem 78(i) this results in $\circ\alpha$ being provable. Thus, the resulting logic would be explosive. ■

In the case of the logic **mbC**, we have called the reader's attention to the fact that the validity of (RP) required the validity of rules (EC) and (EO) (see the end of Subsection 3.2). Interestingly, now in **mCi** we can check that (EC) is enough:

THEOREM 82. In extensions of **mCi** the validity of:

(EC) $\forall\alpha\forall\beta((\alpha \dashv\Vdash \beta)$ implies $(\neg\alpha \dashv\Vdash \neg\beta))$

guarantees the validity of:

(EO) $\forall\alpha\forall\beta((\alpha \dashv\Vdash \beta)$ implies $(\circ\alpha \dashv\Vdash \circ\beta))$.

Proof. Suppose $(\alpha \dashv\Vdash \beta)$. By (EC) we have that $(\neg\alpha \dashv\vdash \neg\beta)$, and from these two equivalences we conclude that $(\alpha \wedge \neg\alpha) \dashv\vdash (\beta \wedge \neg\beta)$. But from Theorems 49(ii) and 76(i) we have that $\neg\circ\gamma \dashv\vdash_{\mathbf{mCi}} (\gamma \wedge \neg\gamma)$, so we have that $\neg\circ\alpha \dashv\vdash \neg\circ\beta$. By Theorem 80(iv) we conclude then that $\circ\alpha \dashv\vdash \circ\beta$. ■

Suppose now we considered the addition to **mCi** of a stronger rule than (EC), in order to ensure replacement:

THEOREM 83. Consider the following rule:

(RC) $\forall\alpha\forall\beta((\alpha \Vdash \beta)$ implies $(\neg\beta \Vdash \neg\alpha))$.

Then, the least extension **L** of **mCi** that satisfies (RC) and proof-by-cases collapses into classical logic.

Proof. From the axioms of **mCi** we first obtain $\neg\circ\alpha \vdash_{\mathbf{L}} \alpha$, and $\neg\circ\alpha \vdash_{\mathbf{L}} \neg\alpha$. By (RC) and Theorem 78(i) we then get $\neg\alpha \vdash_{\mathbf{L}} \circ\alpha$ and $\neg\neg\alpha \vdash_{\mathbf{L}} \circ\alpha$. But then, using proof-by-cases, we conclude that $\vdash_{\mathbf{L}} \circ\alpha$, that is, all formulas are consistent. The result now follows, as usual, from Remark 29. ■

Notice that our paraconsistent formulations of the normal modal logics from Example 34 do *not* extend the logic **mCi** (contrast this with Remark 53 about **mbC**). As we said at the beginning of this subsection, an inconsistency connective $\bullet$ can often be defined from a consistency connective $\circ$ by taking $\sim\circ$, where $\sim$ is a classical negation. The definition of an inconsistency connective by taking $\neg\circ$ is an innovation of **mCi** over **mbC**, and it is typical in fact of most **LFI**s from the literature, as the ones we will be studying in the rest of this chapter. The reader should not think though that the latter class of **C**-systems has any intrinsic advantage over the former. This far, it only seems to have more often met the intuitions of the working paraconsistentists, for some reason or another — or maybe by pure coincidence. At any rate, the distinction between the two classes is only made clear in a framework such as the one set in the present study, where consistency and inconsistency are taken as (primitive or defined) connectives in their own right.

*4.2 Other features of **mCi***

In this subsection we will extend to **mCi** the results obtained for **mbC** in Subsections 3.3, 3.4, 3.5 and 3.6, that is, we will introduce a bivaluation semantics, a possible-translations semantics and a tableau system for **mCi**. Finally, we will exhibit some novel conservative translations from classical logic into **mCi**.

We begin by a brief description of a bivaluation semantics for **mCi**, in the same manner as it was done in Subsection 3.3 with **mbC**. The plan of action is similar to that of **mbC**, and we just outline the main points of departure. First of all, we should remark that, as a consequence of Theorem 121, to be proven at Subsection 5.2, the logic **mCi** cannot be characterized by any finite-valued set of truth-tables, and that gives an extra motivation for the semantics presented in the following.

DEFINITION 84. An **mCi**-valuation is an **mbC**-valuation v: $For^{\circ} \longrightarrow \mathbf{2}$ (see Definition 54) respecting, additionally, the following clauses:

($v6$) $v(\neg\circ\alpha) = 1$ implies $v(\alpha) = 1$ and $v(\neg\alpha) = 1$;
($v7.n$) $v(\circ\neg^{n}\circ\alpha) = 1$ (for $n \geq 0$). ■

The semantic consequence relation obtained from **mCi**-valuations will be denoted by $\models_{\mathbf{mCi}}$. It is easy to prove soundness for **mCi** with respect to **mCi**-valuations.

THEOREM 85. [Soundness] Let $\Gamma\cup\{\alpha\}$ be a set of formulas in For°. Then: $\Gamma \vdash_{\mathbf{mCi}} \alpha$ implies $\Gamma \models_{\mathbf{mCi}} \alpha$. ■

The completeness proof is similar to that of **mbC**, but obviously substituting $\vdash_{\mathbf{mCi}}$ for $\vdash_{\mathbf{mbC}}$. Analogously, given a set of formulas $\Delta\cup\{\alpha\}$ in For° we say that Δ is *relatively maximal with respect to* α *in* **mCi** if $\Delta \nvdash_{\mathbf{mCi}} \alpha$ and for any formula β in For° such that $\beta \notin \Delta$ we have $\Delta, \beta \vdash_{\mathbf{mCi}} \alpha$. As in Lemma 58, relatively maximal theories are closed. An analogue to Lemma 59 can immediately be checked:

LEMMA 86. Let $\Delta \cup \{\alpha\}$ be a set of formulas in For° such that Δ is relatively maximal with respect to α in **mCi**. Then Δ satisfies properties (i)–(v) of Lemma 59, plus the following:

(vi) $\neg\circ\beta \in \Delta$ implies $\beta \in \Delta$ and $\neg\beta \in \Delta$.

(vii) $\circ\neg^{n}\circ\beta \in \Delta$. ■

COROLLARY 87. The characteristic function of a relatively maximal theory of **mCi** defines an **mCi**-valuation. ■

THEOREM 88. [Completeness w.r.t. bivaluation semantics] Let $\Gamma\cup\{\alpha\}$ be a set of formulas in For°. Then $\Gamma \models_{\mathbf{mCi}} \alpha$ implies $\Gamma \vdash_{\mathbf{mCi}} \alpha$. ■

We can obtain a version of Lemma 63 for **mCi**, that is, it is always possible to define an **mCi**-valuation from a given specification of the values of the literals.

LEMMA 89. Let v_0: $\mathcal{P} \cup \{\neg p : p \in \mathcal{P}\} \longrightarrow \mathbf{2}$ be a mapping such that $v_0(\neg p) = 1$ whenever $v_0(p) = 0$ (for $p \in \mathcal{P}$). Then, there exists an **mCi**-valuation v: $For^{\circ} \longrightarrow \mathbf{2}$ extending v_0, that is, such that $v(\varphi) = v_0(\varphi)$ for every $\varphi \in \mathcal{P} \cup \{\neg p : p \in \mathcal{P}\}$.

Proof. The proof is analogous to that of Lemma 63. Thus, we will define

the value of $v(\varphi)$ while doing an induction on the complexity $\ell(\varphi)$ of $\varphi \in For^{\circ}$. Let $v(\varphi) = v_0(\varphi)$ for every $\varphi \in \mathcal{P} \cup \{\neg p : p \in \mathcal{P}\}$, and define $v(p\#q)$ according to clauses $(v1)$–$(v3)$ of Definition 54, for $\# \in \{\wedge, \vee, \rightarrow\}$ and $p, q \in \mathcal{P}$. So, $v(\varphi)$ is defined for every $\varphi \in For^{\circ}$ such that $\ell(\varphi) \leq 1$. Assume that $v(\varphi)$ was defined for every $\varphi \in For^{\circ}$ such that $\ell(\varphi) \leq n$ (for $n \geq 1$) and let $\varphi \in For^{\circ}$ such that $\ell(\varphi) = n + 1$. If $\varphi = (\psi_1 \# \psi_2)$ for $\# \in \{\wedge, \vee, \rightarrow\}$ then $v(\varphi)$ is defined using the corresponding clause from $(v1)$–$(v3)$. If $\varphi = \neg\psi$ then there are two cases to analyze:
(a) $\psi = \neg^k \circ \alpha$, for some $\alpha \in For^{\circ}$ and $k \geq 0$. Then we define $v(\varphi) = 1$ iff $v(\psi) = 0$.
(b) $\psi \neq \neg^k \circ \alpha$, for every $\alpha \in For^{\circ}$ and every $k \geq 0$. Then we define $v(\varphi) = 1$, if $v(\psi) = 0$, and $v(\varphi)$ is defined arbitrarily, otherwise.
Finally, if $\varphi = \circ\psi$ then we set $v(\varphi) = 0$ iff $v(\psi) = v(\neg\psi) = 1$.

It is easy to see that v is an **mCi**-valuation that extends v_0. ■

EXAMPLE 90. With the help of Lemma 89, the bivaluation semantics for **mCi** may be used to show, for instance, that $\neg\neg\alpha \rightarrow \alpha$ is not a thesis of this logic. Indeed, fix $p \in \mathcal{P}$ and consider the mapping

$$v_0 \colon \mathcal{P} \cup \{\neg q : q \in \mathcal{P}\} \longrightarrow \mathbf{2}$$

such that $v(q) = 0$ and $v(\neg q) = 1$ for every $q \in \mathcal{P}$. From the proof of Lemma 89 we know that there exists an **mCi**-valuation $v \colon For^{\circ} \longrightarrow \mathbf{2}$ extending v_0 such that $v(\neg\neg p) = 1$. Then $v(\neg\neg p \rightarrow p) = 0$ and so $\nvdash_{\mathbf{mCi}} (\neg\neg p \rightarrow p)$. By Theorem 85, it follows that $\nvdash_{\mathbf{mCi}} (\neg\neg p \rightarrow p)$. ■

Next, as it was done in Subsection 3.4 with the logic **mbC**, we can also provide an alternative semantics for **mCi** in terms of possible-translations semantics.

Consider the collection $\mathcal{M}_1$ of 3-valued truth-tables formed by the truth-tables of $\mathcal{M}_0$, introduced in Subsection 3.4, but now considering just one consistency operator called $\circ_3$ *instead of* $\circ_1$ and $\circ_2$, presented by the truth-table:

	$\circ_3$
T	T
t	F
F	T

Again, T and t are the designated values. In $\mathcal{M}_1$, the only truth-value that is not consistent is t. If $For_{\mathcal{M}_1}$ denotes the algebra of formulas generated by $\mathcal{P}$ over the signature of $\mathcal{M}_1$, let's consider the set $\textsc{Tr}_1$ of all functions $* \colon For^{\circ} \longrightarrow For_{\mathcal{M}_1}$ respecting the clauses $(tr0)$–$(tr2)$ on translations introduced in Subsection 3.4, plus the following clauses:

$(tr3)_1$ $(\circ\alpha)^* \in \{\circ_3\alpha^*, \circ_3(\neg\alpha)^*\}$;

$(tr3)_2$ if $(\neg\alpha)^* = \neg_1\alpha^*$ then $(\circ\alpha)^* = \circ_3(\neg\alpha)^*$;

$(tr4)_1$ $(\neg^{n+1}\circ\alpha)^* = \neg_1(\neg^n\circ\alpha)^*$.

We say the pair $\mathrm{PT}_1 = \langle \mathcal{M}_1, \mathrm{TR}_1\rangle$ is a *possible-translations semantical structure for* **mCi**. If $\models_{\mathcal{M}_1}$ denotes the consequence relation in $\mathcal{M}_1$, and $\Gamma \cup \{\alpha\}$ is a set of formulas of **mCi**, the PT_1-consequence relation, $\models_{\mathrm{PT}_1}$, is defined as:

$$\Gamma \models_{\mathrm{PT}_1} \alpha \text{ iff } \Gamma^* \models_{\mathcal{M}_1} \alpha^* \text{ for all } * \in \mathrm{TR}_1.$$

We leave to the reader the proof of the following easy result:

THEOREM 91. [Soundness] Let $\Gamma\cup\{\alpha\}$ be a set of formulas of **mCi**. Then $\Gamma \vdash_{\mathbf{mCi}} \alpha$ implies $\Gamma \models_{\mathrm{PT}_1} \alpha$. ■

The completeness proof follows the same lines than the one obtained for **mbC** (cf. [Marcos, 2005f]).

THEOREM 92. [Representability] Given an **mCi**-valuation v there is a translation $*$ in TR_1 and a valuation w in $\mathcal{M}_1$ such that, for every formula α in **mCi**:

$$w(\alpha^*) = T \text{ implies } v(\neg\alpha) = 0; \text{ and}$$

$$w(\alpha^*) = F \text{ iff } v(\alpha) = 0.$$

Proof. The proof is similar to that of Theorem 67, but now defining $(\circ\alpha)^* = \circ_3(\neg\alpha)^*$ if $v(\neg\alpha) = 0$, and $(\circ\alpha)^* = \circ_3\alpha^*$ otherwise. Finally, set $(\neg^{n+1}\circ\alpha)^* = \neg_1(\neg^n\circ\alpha)^*$. Details are left to the reader. ■

COROLLARY 93. [Completeness w.r.t. possible-translations semantics] Let $\Gamma\cup\{\alpha\}$ be a set of formulas in **mCi**. Then $\Gamma \models_{\mathrm{PT}_1} \alpha$ implies $\Gamma \vdash_{\mathbf{mCi}} \alpha$.

In Remark 43 we have defined two supplementing negations for **mbC**, $\wr$ and $\sim$, and in Remark 70 we have shown that only one of them, namely $\sim$, was classical in **mbC**. Now we can use the possible-translations semantics of **mCi** to check that in this logic the two negations produce equivalent formulas:

THEOREM 94. Given a formula α, the formulas $\wr\alpha$ and $\sim\alpha$ are equivalent in **mCi**. As a result, $\wr$ defines a classical negation in **mCi**.

Proof. Notice that, using the above possible-translations semantics for **mCi**, the formulas $\wr p$ and $\sim p$ produce exactly the same truth-tables. ■

Now, using the general techniques introduced in [Caleiro *et al.*, 2005b] we can easily obtain an adequate tableau system for **mCi**, in the same way

that was done for **mbC** in Subsection 3.5. Thus, in view of the bivaluation semantics for **mCi** stated in Definition 84 from the bivaluation semantics for **mbC**, it is enough to define the following:

DEFINITION 95. We define a tableau system for **mCi** by adding to the tableau system for **mbC** introduced in Example 71 the following rules:

$$\frac{T(\neg\circ X)}{T(X),\ T(\neg X)} \qquad \frac{}{T(\circ\,\neg^n\circ X)} \quad (\text{for } n \geq 0) \qquad \blacksquare$$

Finally, let's talk again about classical logic. In Theorem 74 of Subsection 3.6 we have seen how **CPL** can be encoded inside **mbC** through a conservative translation. Clearly, that same translation works for **mCi**. We will now show how it is possible to encode **eCPL** inside **mCi**, in a similar fashion.

THEOREM 96. Let For° be the algebra of formulas for the signature Σ° of **mCi**. Consider any mapping t_2: $For^\circ \longrightarrow For^\circ$ such that:

1. $t_2(p) = p$, for every $p \in \mathcal{P}$;
2. $t_2(\gamma \# \delta) = t_2(\gamma) \# t_2(\delta)$, if $\# \in \{\wedge, \vee, \rightarrow\}$;
3. $t_2(\neg\gamma) \in \{\sim t_2(\gamma),\ \wr\, t_2(\gamma)\}$;
4. $t_2(\circ\gamma) = \circ\circ t_2(\gamma)$.

Then, t_2 is a conservative translation from **eCPL** to **mCi**.

Proof. The proof is almost identical to that of Theorem 74. The only novel clause is 4, but it is clear how it works (recall axiom $(\text{cc})_0$). $\blacksquare$

4.3 Inconsistency operator as primitive

Up to now we have concentrated almost exclusively on the formal notion of consistency; formal inconsistency has appeared only derivatively, defined with the help of a classical or of a paraconsistent negation. It is equally natural, however, to provide alternative axiomatizations for the logic **mCi** or its close relatives starting from a primitive inconsistency connective. We will now show how to do this in two different ways, one in terms of $\circ$ and $\bullet$, and the other in terms of $\bullet$ alone.

Let $\Sigma^\bullet$ and $\Sigma^{\circ\bullet}$ be the extensions of the signature Σ (recall Remark 15) obtained by the addition, respectively, of a new unary connective $\bullet$ and of two unary connectives $\circ$ and $\bullet$. Let $For^\bullet$ and $For^{\circ\bullet}$ be the respective algebras of formulas. The idea of axiomatizing **mCi** just in terms of $\bullet$ involves the assumption that $\bullet\alpha$ means $\neg\circ\alpha$ while $\neg\bullet\alpha$ means $\circ\alpha$. As a consequence, axiom schemas (bc1), (ci) and $(\text{ci})_n$ should adopt the following forms:

(bc1)′ $\neg\bullet\alpha \rightarrow (\alpha \rightarrow (\neg\alpha \rightarrow \beta))$

(ci)′ $\bullet\alpha \rightarrow (\alpha \wedge \neg\alpha)$

(cc)′$_n$ $\neg\bullet\neg^n\bullet\alpha \qquad (n \geq 0)$

This leads to the following definition:

DEFINITION 97. The logic **mCi$^\bullet$** defined over signature $\Sigma^\bullet$ is defined by the axiom schemas (Ax1)–(Ax10) (recall Definition 28) plus the axiom schemas (bc1)′, (ci)′ and (cc)′$_n$ (for $n \geq 0$) introduced above, together with (MP). ■

The next result will demonstrate to which extent **mCi** and **mCi$^\bullet$** are 'the same logic'. Since these logics are written in distinct signatures, an appropriate way of comparing them is by way of (some very strict and specific) translations.

THEOREM 98.
(i) Let $+ : For^\circ \longrightarrow For^\bullet$ be a mapping defined as follows:

1. $p^+ = p$ if $p \in \mathcal{P}$;
2. $(\alpha \# \beta)^+ = (\alpha^+ \# \beta^+)$ where $\# \in \{\wedge, \vee, \rightarrow\}$;
3. $(\neg\circ\alpha)^+ = \bullet\alpha^+$;
4. $(\neg\alpha)^+ = \neg\alpha^+$ if $\alpha \neq \circ\beta$ for every β;
5. $(\circ\alpha)^+ = \neg\bullet\alpha^+$.

Then, the mapping $+$ is a translation from **mCi** to **mCi$^\bullet$**, that is, for every $\Gamma \cup \{\alpha\} \subseteq For^\circ$:

$$\Gamma \vdash_{\mathbf{mCi}} \alpha \text{ implies } \Gamma^+ \vdash_{\mathbf{mCi}^\bullet} \alpha^+.$$

(ii) Let $- : For^\bullet \longrightarrow For^\circ$ be a mapping defined as follows:

1. $p^- = p$ if $p \in \mathcal{P}$;
2. $(\alpha \# \beta)^- = (\alpha^- \# \beta^-)$ where $\# \in \{\wedge, \vee, \rightarrow\}$;
3. $(\neg\bullet\alpha)^- = \circ\alpha^-$;
4. $(\neg\alpha)^- = \neg\alpha^-$ if $\alpha \neq \bullet\beta$ for every β;
5. $(\bullet\alpha)^- = \neg\circ\alpha^-$.

Then, the mapping $-$ is a translation from **mCi$^\bullet$** to **mCi**, that is, for every $\Gamma \cup \{\alpha\} \subseteq For^\bullet$:

$$\Gamma \vdash_{\mathbf{mCi}^\bullet} \alpha \text{ implies } \Gamma^- \vdash_{\mathbf{mCi}} \alpha^-.$$

Proof. Item (i). Suppose that $\Gamma \vdash_{\mathbf{mCi}} \alpha$. By induction on the length of a derivation in **mCi** of α from Γ, it can be easily proven that $\Gamma^{+} \vdash_{\mathbf{mCi}^{\bullet}} \alpha^{+}$. There are just three cases that deserve some attention. These cases occur when $\alpha \in For^{\circ}$ is an instance of an axiom in **mCi** but α^{+} is not an instance of an axiom in $\mathbf{mCi}^{\bullet}$. These three cases are: (1) $\alpha = \circ\circ\gamma \rightarrow (\circ\gamma \rightarrow (\neg\circ\gamma \rightarrow \delta))$ for $\gamma, \delta \in For^{\circ}$; (2) $\alpha = (\circ\gamma \vee \neg\circ\gamma)$; and (3) $\alpha = \neg\circ\circ\gamma \rightarrow (\circ\gamma \wedge \neg\circ\gamma)$. In case (1), $\alpha^{+} = \neg\bullet\neg\bullet\gamma^{+} \rightarrow (\neg\bullet\gamma^{+} \rightarrow (\bullet\gamma^{+} \rightarrow \delta^{+}))$, which is not an instance of an axiom in $\mathbf{mCi}^{\bullet}$. However, it is immediate to see that α^{+} is a theorem of $\mathbf{mCi}^{\bullet}$, because of axioms (bc1)′, (ci)′ and by the deduction theorem. Indeed, $\bullet\gamma^{+} \vdash_{\mathbf{mCi}^{\bullet}} (\gamma^{+} \wedge \neg\gamma^{+})$ and $\neg\bullet\gamma^{+}, (\gamma^{+} \wedge \neg\gamma^{+}) \vdash_{\mathbf{mCi}^{\bullet}} \delta^{+}$, therefore $\neg\bullet\neg\bullet\gamma^{+}, \neg\bullet\gamma^{+}, \bullet\gamma^{+} \vdash_{\mathbf{mCi}^{\bullet}} \delta^{+}$. Using the deduction theorem it then follows that $\vdash_{\mathbf{mCi}^{\bullet}} \alpha^{+}$. In case (2), $\alpha^{+} = (\neg\bullet\gamma^{+} \vee \bullet\gamma^{+})$, which is not an axiom of $\mathbf{mCi}^{\bullet}$, yet it can be easily checked to be a theorem of $\mathbf{mCi}^{\bullet}$. In case (3), $\alpha^{+} = \bullet\neg\bullet\gamma^{+} \rightarrow (\neg\bullet\gamma^{+} \wedge \bullet\gamma^{+})$. This is not an axiom, but it can be easily proven in $\mathbf{mCi}^{\bullet}$. Indeed, by $(\mathrm{cc})'_1$, (bc1)′, the deduction theorem and proof-by-cases it follows that $\neg\neg\bullet\delta \vdash_{\mathbf{mCi}^{\bullet}} \bullet\delta$ holds in $\mathbf{mCi}^{\bullet}$, for every δ. Using this, (ci)′, properties of the standard conjunction and the deduction theorem, it follows that α^{+} is a theorem of $\mathbf{mCi}^{\bullet}$.
Item (ii). The proof is entirely analogous to that of item (i). ■

The fact that both logics are inter-translatable means that **mCi** encodes $\mathbf{mCi}^{\bullet}$ and vice-versa. Moreover, we could take the combined logic $\mathbf{mCi}^{\circ\bullet}$ defined over $\Sigma^{\circ\bullet}$ by putting together all the axiom schemas of **mCi** and $\mathbf{mCi}^{\bullet}$, plus (MP) (technically, $\mathbf{mCi}^{\circ\bullet}$ can be obtained as the fibring of **mCi** and $\mathbf{mCi}^{\bullet}$; see, for instance, the entry on fibring [Caleiro *et al.*, 2005] in this Handbook). It is also possible to show that the logic $\mathbf{mCi}^{\circ\bullet}$ is a conservative extension of both **mCi** and $\mathbf{mCi}^{\bullet}$. The following result is easy to check:

THEOREM 99. Let α be a formula in $For^{\circ\bullet}$. Then

$$\circ\alpha \dashv\vdash_{\mathbf{mCi}^{\circ\bullet}} \neg\bullet\alpha \quad \text{and} \quad \neg\circ\alpha \dashv\vdash_{\mathbf{mCi}^{\circ\bullet}} \bullet\alpha.$$ ■

However, as yet another witness to the fact that the replacement property (RP) (see Remark 51) is not enjoyed by these logics, it is not difficult to see (say, by means of bivaluations) that, in general, the following is true, for $\alpha \in For^{\circ}$ and $\beta \in For^{\bullet}$:

$$\alpha \nvdash_{\mathbf{mCi}} (\alpha^{+})^{-}, \quad (\alpha^{+})^{-} \nvdash_{\mathbf{mCi}} \alpha, \quad \alpha \nvdash_{\mathbf{mCi}^{\circ\bullet}} \alpha^{+}, \quad \alpha^{+} \nvdash_{\mathbf{mCi}^{\circ\bullet}} \alpha$$

$$\beta \nvdash_{\mathbf{mCi}^{\bullet}} (\beta^{-})^{+}, \quad (\beta^{-})^{+} \nvdash_{\mathbf{mCi}^{\bullet}} \beta, \quad \beta \nvdash_{\mathbf{mCi}^{\circ\bullet}} \beta^{-}, \quad \beta^{-} \nvdash_{\mathbf{mCi}^{\circ\bullet}} \beta.$$

The corresponding bivaluation semantics, possible-translations semantics and tableau procedures for the versions of **mCi** in the above signatures can be easily implemented and we will not annoy the reader with details.

*4.4 Enhancing **mCi** in dealing with double negations*

In this subsection we will see what happens when the logics **mbC** and **mCi** are further extended with axioms dealing with doubly negated formulas, namely:

(cf) $\neg\neg\alpha \rightarrow \alpha$
(ce) $\alpha \rightarrow \neg\neg\alpha$

Note that (cf) has already appeared as (Ax11) in Definition 28. From item (iii) of Theorem 77 we know that neither (ce) nor (cf) is provable in **mCi**. Adding such axioms makes the negation of this logic a bit closer to classical negation. Moreover, we will see that adding them helps in simplifying the axiomatic presentations of the resulting logics, and it also has a nice consequence for the interaction of negation with the connectives for consistency and inconsistency.

DEFINITION 100. Consider the signature Σ°. Recall the axiomatizations of **mbC** and **mCi** from Definitions 42 and 75. Then:

1. **bC** is axiomatized as **mbC** plus (cf).
2. **Ci** is axiomatized as **mCi** plus (cf).
3. **mbCe** is axiomatized as **mbC** plus (ce).
4. **mCie** is axiomatized as **mCi** plus (ce).
5. **bCe** is axiomatized as **bC** plus (ce).
6. **Cie** is axiomatized as **Ci** plus (ce). ■

It is easy to check that:

THEOREM 101.
(i) $\circ\alpha \vdash_{\mathbf{Ci}} \circ\neg\alpha$;
(ii) $\bullet\neg\alpha \vdash_{\mathbf{Ci}} \bullet\alpha$;
(iii) $\circ\neg\alpha \vdash_{\mathbf{mCie}} \circ\alpha$;
(iv) $\bullet\alpha \vdash_{\mathbf{mCie}} \bullet\neg\alpha$. ■

Using the latter result one might provide a simpler and finitary axiomatization for the logic **Ci** (thus also for **Cie**):

THEOREM 102. The logic **Ci** may be obtained from **mbC** by adding the axiom schemas (ci) (see Definition 75) and (cf) (Subsection 4.4), to wit:

(ci) $\neg\circ\alpha \rightarrow (\alpha \wedge \neg\alpha)$
(cf) $\neg\neg\alpha \rightarrow \alpha$

Proof. Let $\Vdash$ be the consequence relation of the logic obtained from **mbC** by adding the axiom schemas (ci) and (cf). Of course $\Vdash \subseteq \vdash_{\mathbf{Ci}}$. In order to prove the converse, it is enough to prove that $\Vdash \circ\neg^{n}\circ\alpha$ (that is, axiom schema $(\mathrm{cc})_n$) holds good for every formula α and every natural number n.

For $n = 0$ note that $\circ\alpha, \alpha, \neg\alpha \Vdash \circ\circ\alpha$, by (bc1), and $\neg\circ\alpha \Vdash \alpha \wedge \neg\alpha$, by (ci). In particular $\neg\circ\circ\alpha \Vdash \circ\alpha \wedge \neg\circ\alpha$, thus $\neg\circ\circ\alpha \Vdash \circ\circ\alpha$. But $\circ\circ\alpha \Vdash \circ\circ\alpha$ and then proof-by-cases gives us

$$(*) \qquad \Vdash \circ\circ\alpha.$$

Now, by (ci) again, we have that $\neg\circ\neg\alpha \Vdash \neg\alpha \wedge \neg\neg\alpha$ and then $\neg\circ\neg\alpha \Vdash \neg\alpha \wedge \alpha$, by (cf). Using (bc1) we obtain $\circ\alpha, \alpha, \neg\alpha \Vdash \circ\neg\alpha$ and so $\circ\alpha, \neg\circ\neg\alpha \Vdash \circ\neg\alpha$. Since $\circ\alpha, \circ\neg\alpha \Vdash \circ\neg\alpha$ then proof-by-cases gives us $\circ\alpha \Vdash \circ\neg\alpha$ for every α, as in Theorem 101(i). In particular,

$$(**) \qquad \circ\neg^{n}\circ\alpha \Vdash \circ\neg^{n+1}\circ\alpha$$

for every $n \geq 0$ and every α. Using $(*)$ and $(**)$, it is now immediate to obtain $(\mathrm{cc})_n$ by induction on n. ■

As regards semantic presentations, in view of Theorem 121 (see Subsection 5.2), we know that the logics from Definition 100 are not characterizable by a collection of finite-valued truth-tables. However, it is straightforward to endow these new systems with adequate bivaluation semantics, using the methods from previous sections. Indeed:

THEOREM 103. Axiom (cf) corresponds to the following clause on the definition of a bivaluation semantics:

$(v8)$ $v(\neg\neg\alpha) = 1$ implies $v(\alpha) = 1$.

Similarly, axiom (ce) corresponds to:

$(v9)$ $v(\alpha) = 1$ implies $v(\neg\neg\alpha) = 1$. ■

Accordingly, one can now prove, for instance, that **Ci** is sound and complete for the class of bivaluations v: $For^{\circ} \longrightarrow \mathbf{2}$ satisfying clauses $(v1)$–$(v5)$ of Definition 54 plus clause $(v6)$ of Definition 84 and clause $(v8)$ of Theorem 103.

The next useful result concerning the definability of bivaluations for the systems introduced in Definition 100 can be obtained. The proof is done by appropriately adapting the proofs of Lemmas 63 and 89.

LEMMA 104. Let v_0: $\mathcal{P} \cup \{\neg p : p \in \mathcal{P}\} \longrightarrow \mathbf{2}$ be a mapping such that $v_0(\neg p) = 1$ whenever $v_0(p) = 0$ (for $p \in \mathcal{P}$). Then, there exist bivaluations extending v_0, for each one of the logics introduced in Definition 100.

Proof. We only prove the case for **Ci**. Thus, given v_0, define $v(\varphi) = v_0(\varphi)$ for every $\varphi \in \mathcal{P} \cup \{\neg p : p \in \mathcal{P}\}$, and $v(p\#q)$ is defined according to clauses $(v1)$–$(v3)$ of Definition 54, for $\# \in \{\wedge, \vee, \rightarrow\}$ and $p, q \in \mathcal{P}$. Suppose that $v(\varphi)$ was defined for every $\varphi \in For^{\circ}$ such that $\ell(\varphi) \leq n$ (for $n \geq 1$) and let $\varphi \in For^{\circ}$ such that $\ell(\varphi) = n + 1$. If $\varphi = (\psi_1 \# \psi_2)$ for $\# \in \{\wedge, \vee, \rightarrow\}$ then we use clauses $(v1)$–$(v3)$ to define $v(\varphi)$. If $\varphi = \circ\psi$ then define $v(\varphi) = 0$ iff $v(\psi) = v(\neg\psi) = 1$.

Finally, suppose that $\varphi = \neg\psi$. If $v(\psi) = 0$ then define $v(\varphi) = 1$. On the

other hand, if $v(\psi) = 1$ then there are three cases to analyze:
(a) $\psi = \circ\alpha$, for some $\alpha \in For^\circ$. Then, define $v(\varphi) = 0$.
(b) $\psi = \neg\alpha$, for some $\alpha \in For^\circ$, such that $v(\alpha) = 0$. Then we define $v(\varphi) = 0$.
(c) In any other case, $v(\varphi)$ is defined arbitrarily.
It is straightforward to check that v is a **Ci**-valuation extending v_0. The proof for the other systems is entirely analogous, and we leave the details to the reader. ■

We can also obtain adequate tableaux for these systems, as in previous sections. Possible-translations semantics for **bC**, **Ci**, **bCe** and **Cie** may be found in [Marcos, 2005f]. These four logics were exhaustively studied in [Carnielli and Marcos, 2002]. Non-deterministic semantics for these logics can be found in [Avron, 2005a].

5 ADDITIONAL TOPICS ON **LFI**S

*5.1 The **dC**-systems*

As we have seen in Theorem 98, the formulas $\bullet\alpha$ and $\neg\circ\alpha$ have the same meaning (up to translations) in **mCi**. Moreover, we also know from Theorem 49(i) and axiom (ci) that the formulas $\bullet\alpha$ and $(\alpha \wedge \neg\alpha)$ are equivalent in **mCi**. However, as we know from Theorem 76, the formulas $\neg\bullet\alpha$ and $\neg(\alpha \wedge \neg\alpha)$ are not equivalent, nor are the formulas $\neg\neg\bullet\alpha$ and $\neg\neg(\alpha \wedge \neg\alpha)$, and so on.

It seemed only natural, thus, to consider extensions of **mCi** in which the meaning of statements involving $\bullet$ (and also $\circ$) may be recast in terms of the other connectives, by means of translations or of explicit definitions. This maneuver led us to the class of **LFI**s known as **dC**-systems, in which the new connective of consistency may be dismissed from the beginning, and replaced by a formula built from the other connectives already present in the signature (recall Definition 32).[8] The logic **Cil**, to be defined below, is an example of this strategy.

DEFINITION 105. The logic **Cil**, defined over the signature Σ°, is obtained from **Ci** by the addition of the following axiom schema:

(cl) $\neg(\alpha \wedge \neg\alpha) \rightarrow \circ\alpha$

Other logics may be obtained in a similar fashion, such as the logic **Cile**, defined by the addition of (ce) to **Cil** (recall Subsection 4.4). ■

[8]The reader is invited to adapt Definition 32 to deal also with the inconsistency operator, and to logics defined over signatures $\Sigma^\bullet$ and $\Sigma^{\circ\bullet}$.

By the very definition of **Cil**, it is clear there cannot be a paraconsistent extension of **Cil** in which the schema $\neg(\alpha \wedge \neg\alpha)$ is provable. There are, however, other paraconsistent extensions of **Ci**, such as **LFI1** (see Example 18 and Theorem 127) or extensions of **bC** such as all non-degenerate normal modal logics extending the system KT (recall Example 34), in which the schema $\neg(\alpha \wedge \neg\alpha)$ is indeed provable.

It can be checked that **Cil** is in fact an indirect a **dC**-system based on classical logic:

THEOREM 106. The logic **Cil** may be defined over Σ by identifying $\circ\alpha$ with $\neg(\alpha \wedge \neg\alpha)$. More precisely: Let $\mathbf{Cil}^{\copyright}$ be the logic over Σ defined by axiom schemas (Ax1)–(Ax11) (see Definition 28), rule (MP), plus the following axiom schema:

(bc1)″ $\neg(\alpha \wedge \neg\alpha) \rightarrow (\alpha \rightarrow (\neg\alpha \rightarrow \beta))$

Let $\star : For^{\circ} \longrightarrow For$ be a mapping defined as follows:

1. $p^{\star} = p$ if $p \in \mathcal{P}$;
2. $(\alpha \# \beta)^{\star} = (\alpha^{\star} \# \beta^{\star})$ where $\# \in \{\wedge, \vee, \rightarrow\}$;
3. $(\neg\alpha)^{\star} = \neg\alpha^{\star}$;
4. $(\circ\alpha)^{\star} = \neg(\alpha^{\star} \wedge \neg\alpha^{\star})$.

Then, the mapping $\star$ is a translation from **Cil** to $\mathbf{Cil}^{\copyright}$, that is, for every $\Gamma \cup \{\alpha\} \subseteq For^{\circ}$:

$$\Gamma \vdash_{\mathbf{Cil}} \alpha \text{ implies } \Gamma^{\star} \vdash_{\mathbf{Cil}^{\copyright}} \alpha^{\star}.$$

On the other hand, **Cil** is a conservative extension of $\mathbf{Cil}^{\copyright}$, that is, for every $\Gamma \cup \{\alpha\} \subseteq For$:

$$\Gamma \vdash_{\mathbf{Cil}^{\copyright}} \alpha \text{ iff } \Gamma \vdash_{\mathbf{Cil}} \alpha.$$

As a consequence of the above, the following holds good, for every $\Gamma \cup \{\alpha\} \subseteq For^{\circ}$:

$$\Gamma \vdash_{\mathbf{Cil}} \alpha \text{ implies } \Gamma^{\star} \vdash_{\mathbf{Cil}} \alpha^{\star}.$$

Proof. The proof follows the lines of the proof of Theorem 98, and there is just one further critical case to analyze: Any axiom of **Cil** of the form $\alpha = \neg(\gamma \wedge \neg\gamma) \rightarrow \circ\gamma$ is translated as $\alpha^{\star} = \neg(\gamma^{\star} \wedge \neg\gamma^{\star}) \rightarrow \neg(\gamma^{\star} \wedge \neg\gamma^{\star})$, which is not an axiom of $\mathbf{Cil}^{\copyright}$, but it is obviously a theorem of $\mathbf{Cil}^{\copyright}$. This shows that $\star$ is a translation from **Cil** to $\mathbf{Cil}^{\copyright}$.

Consider now a set $\Gamma \cup \{\alpha\} \subseteq For$. Observe that every axiom of $\mathbf{Cil}^{\copyright}$ different from (bc1)″ is an axiom of **Cil**. On the other hand, it is easy to see (using the deduction theorem) that (bc1)″ is a theorem of **Cil**. Hence, by induction on the length of a derivation in $\mathbf{Cil}^{\copyright}$ of α from Γ it follows

that, if $\Gamma \vdash_{\mathbf{Cil}^{\copyright}} \alpha$ then $\Gamma \vdash_{\mathbf{Cil}} \alpha$. Conversely, if $\Gamma \vdash_{\mathbf{Cil}} \alpha$ then $\Gamma^{\star} \vdash_{\mathbf{Cil}^{\copyright}} \alpha^{\star}$, since $\star$ is a translation from **Cil** to $\mathbf{Cil}^{\copyright}$. But, if $\beta \in For$ then $\beta^{\star} = \beta$, so $\Gamma \vdash_{\mathbf{Cil}^{\copyright}} \alpha$. This shows that **Cil** is a conservative extension of $\mathbf{Cil}^{\copyright}$. The rest of the proof is straightforward. ■

The above theorem shows that $\mathbf{Cil}^{\copyright}$ is a direct **dC**-system, just as much as **Cil** is an indirect one (recall Definitions 32 and 33). Observe that (bc1)″ was already introduced in Definition 28 as axiom (bc1). Thus, the logic $\mathbf{Cil}^{\copyright}$ is obtained from C_1 by the elimination of axioms (ca1)–(ca3) (or, equivalently, C_1 is obtained from $\mathbf{Cil}^{\copyright}$ by adding axiom schemas (ca1)–(ca3); see Definition 108 and Remark 109). The formula schema $\neg(\alpha \wedge \neg\alpha)$ played an important role in the original construction of the logics C_n, and it has often been identified with the so-called 'Principle of Non-Contradiction'. Notice, however, that such an identification is not possible with our present definition of this principle (Principle (1) in Subsection 2.1).

There is no consensus in the literature on what concerns the status of the schema $\neg(\alpha \wedge \neg\alpha)$ inside paraconsistent logics. Its validity has been criticized by some (see, for instance, [Béziau, 2002a]). A good technical reason for expecting this schema to fail is connected to the possible consequent failure of the replacement property, as predicted in Theorem 52(iv). On the other hand, the proposal of paraconsistent logics in which this schema does not hold has *also* been criticized, as for instance in [Routley and Meyer, 1976], where the authors claim that, for dialectical logics (i.e. for logics disrespecting our version of the Principle of Non-Contradiction), not only do we usually have that $\neg(\alpha \wedge \neg\alpha)$ is a theorem, but that feature does not conflict with other logical truths of such logics. On our approach, the whole controversy seems artificial and ill-advised. It might well be just a sterile offspring of the misidentification of the Principle of Explosion and the Principle of Non-Contradiction: In general, only the former should worry a paraconsistent logician, the latter being a much less demanding and a very often strictly observed principle (check the ensuing discussion in section 3.8 of [Carnielli and Marcos, 2002]).

Using (bc1) and (cl), every theorem of the form $\circ(\alpha \wedge \neg\alpha)$ can be proven. In the presence of axiom (cf), as in Theorem 101(i), this allows one to prove, in **Cil**, every theorem of the form $\circ\neg^{n}(\alpha \wedge \neg\alpha)$. This feature was to raise protests by some authors (see for instance [Sylvan, 1990]), according to whom it makes no sense to declare contradictions (case $n = 0$ in the above formula) to be provably consistent.

With respect to semantics, Theorem 125 (see Subsection 5.2) proves that the logics **Cil** and $\mathbf{Cil}^{\copyright}$ are not characterizable by a collection of finite-valued truth-tables. Of course, we can obtain a bivaluation semantics for $\mathbf{Cil}^{\copyright}$ by considering mappings v: $For \longrightarrow \mathbf{2}$ satisfying axioms (v1)–(v4) of Definition 54, plus the following:

(v10) $v(\neg(\alpha \wedge \neg\alpha)) = 1$ implies $v(\alpha) = 0$ or $v(\neg\alpha) = 0$;

$(v11)$ $v(\neg\neg\alpha) = 1$ implies $v(\alpha) = 1$.

In the case of **Cil**, one may consider bivaluations $v\colon For^{\circ} \longrightarrow \mathbf{2}$ that satisfy axioms $(v1)$–$(v5)$ of Definition 54, plus $(v6)$ (see Definition 84) and $(v11)$. Of course, a result analogous to Lemma 104 can be stated and proven for the logics **Cil** and **Cil**$^{\copyright}$. At this point it should be obvious to the reader how the tableaux for these logics would look like.

If the reader has still not gotten used to the frequent failure of the replacement property, he might be surprised with the following asymmetry allowed by the logic **Cil**. The consistency operator $\circ\alpha$ is equivalent in **Cil** to the formula $\neg(\alpha \wedge \neg\alpha)$ (cf. Definition 105) and consequently the logic resulting from the addition of $\neg(\alpha \wedge \neg\alpha)$ to **Cil** is no longer paraconsistent. On the other hand:

THEOREM 107. The logic resulting from the addition of $\neg(\neg\alpha \wedge \alpha)$ to **Cil** is still paraconsistent, and so the operator $\circ$ cannot be alternatively expressed by the formula $\neg(\neg\alpha \wedge \alpha)$.

Proof. The first collection of truth-tables from the proof of Theorem 50 provides a model of **Cil** plus $\neg(\neg\alpha \wedge \alpha)$. The same collection of truth-tables show that there are atomic formulas p and q such that $\neg(\neg p \wedge p), p, \neg p$ take designated values, while q does not: Just assign the value $\frac{1}{2}$ to p and 0 to q. ■

The above asymmetry has been sharply pointed out in Theorem 4 of [Urbas, 1989] for the case of the logic C_1 which is, as we mentioned before, an extension of **Cil**$^{\copyright}$ (see also Remark 109). This asymmetry remained hidden for a long time within the realm of the logics C_n. Indeed, the first decision procedure offered for the logic C_1 in terms of quasi matrices, in [da Costa and Alves, 1977], was mistaken exactly in assuming $\neg(\alpha \wedge \neg\alpha)$ and $\neg(\neg\alpha \wedge \alpha)$ to be equivalent formulas.

Some natural alternatives to (cl) can immediately be considered:

(cd) $\neg(\neg\alpha \wedge \alpha) \rightarrow \circ\alpha$;

(cb) $(\neg(\alpha \wedge \neg\alpha) \vee \neg(\neg\alpha \wedge \alpha)) \rightarrow \circ\alpha$.

(RG) $\beta \dashv\vdash \alpha \wedge \neg\alpha$ implies $\neg\beta \dashv\vdash \neg(\alpha \wedge \neg\alpha)$

Clearly, the addition to **Ci** of the axiom (cd) instead of the axiom (cl), would produce a logic in which the asymmetry pointed out in Theorem 107 is inverted. That inconvenient can be solved if the axiom (cb) is added instead, as that move produces a logic in which both $\neg(\alpha\wedge\neg\alpha)$ and $\neg(\neg\alpha\wedge\alpha)$ express consistency. However, that will not make the difficulties about the replacement property, (RP), go away. In fact, the equivalence of similar more complex formulas would not be guaranteed by (cb): It can be shown for instance that formulas such as $\neg(\alpha \wedge (\alpha \wedge \neg\alpha))$ and $\neg((\alpha \wedge \neg\alpha) \wedge \alpha)$ are

not automatically equivalent, even though $(\alpha \wedge (\alpha \wedge \neg\alpha))$ and $((\alpha \wedge \neg\alpha) \wedge \alpha)$ are equivalent on any **C**-system based on (positive) classical logic. As pointed out in [Carnielli and Marcos, 2002], a way of solving that specific predicament without necessarily going as far as establishing the validity of (RP) is simply by adding the rule (RG). Of course, **dC**-systems with full (RP) are clearly available, as it has been illustrated by the modal logics proposed in Example 34, all of which extend the fundamental **C**-system **mbC** (recall Remark 53).

It should be clear that each **dC**-system can in principle generate an infinite number of other **dC**-systems, if one applies to it the same strategy as that of the C_n logics, for $1 \leq n < \omega$ (cf. Definition 28), namely, if one simply requires stronger and stronger conditions to be met in order to establish the consistency of a formula.

5.2 Adding modularity: Letting consistency propagate

Given a class of consistent formulas, an important issue is to understand how this consistency propagates towards simpler or more complex formulas. As we have seen in Theorem 101, the addition to **mCi** of axioms or rules controlling the behavior of doubly negated formulas reflects directly on the propagation of consistency through negation. As we will see in this subsection, one can in fact produce interesting variations on the recipe that constructs **LFI**s by directly controlling the way consistency propagates.

DEFINITION 108.
(i) The logic **Cia** is obtained by the addition of the following axiom schemas to **Ci** (see Definition 100):

(ca1) $(\circ\alpha \wedge \circ\beta) \rightarrow \circ(\alpha \wedge \beta)$;
(ca2) $(\circ\alpha \wedge \circ\beta) \rightarrow \circ(\alpha \vee \beta)$;
(ca3) $(\circ\alpha \wedge \circ\beta) \rightarrow \circ(\alpha \rightarrow \beta)$.

(ii) The logic **Cila** is obtained by the addition of the axiom schema (cl) to **Cia** or, equivalently, of the axioms (ca1)–(ca3) to **Cil** (see Definition 105). Using axioms (cd) or (cb) instead of (cl) one might similarly define the logics **Cida** or **Ciba**. Adding axiom (ce) to those systems one might define the logics **Cilae**, **Cidae** and **Cibae**. ■

REMARK 109. It is worth insisting that the only difference between **Cila** and the original formulation of C_1 (recall Definition 28) is that the connective $\circ$ in C_1 was not taken as primitive, but $\circ\alpha$, originally denoted as α°, was assumed from the start to be an abbreviation of the formula $\neg(\alpha \wedge \neg\alpha)$. A transformation to that same effect is done by the translation $\star$ from Theorem 106. However, it should be noted that there are formulas $\alpha \in For^\circ$ such that α and $\alpha^\star$ are not equivalent in **Cila**. On the other hand, C_1 coincides with $\mathbf{Cila}^{\copyright}$, the logic obtained from $\mathbf{Cil}^{\copyright}$ (see again Theorem 106)

by adding axioms (ca1)–(ca3). In other words, **Cila** is obtained from C_1 by adding the consistency operator $\circ$ to the signature as well as the obvious axioms stating the equivalence between the formulas $\circ\alpha$ and $\neg(\alpha \wedge \neg\alpha)$. In the terminology of Definition 33, we may say that **Cila** *corresponds* to C_1. For the other logics in the hierarchy C_n, $1 \leq n < \omega$, the formula $\circ\alpha$ abbreviates more and more complex formulas, or sets of formulas, as it can be seen in Definition 28. ■

The logics **Cila** and C_1 are not exactly coincident since they are defined over distinct signatures. However, they are related by means of translations in the same way as **Cil** and **Cil**$^{\copyright}$ were so related (recall Theorem 106). In other terms, **Cila** is an indirect **dC**-system, while C_1 is a direct **dC**-system, as the theorem below shows.

THEOREM 110. Let $\star : For^{\circ} \longrightarrow For$ be the translation mapping defined as in Theorem 106. Then $\star$ is a translation from **Cila** to C_1, that is, for every $\Gamma \cup \{\alpha\} \subseteq For^{\circ}$,

$$\Gamma \vdash_{\mathbf{Cila}} \alpha \text{ implies } \Gamma^{\star} \vdash_{C_1} \alpha^{\star}.$$

On the other hand, **Cila** is a conservative extension of C_1, that is, for every $\Gamma \cup \{\alpha\} \subseteq For$,

$$\Gamma \vdash_{C_1} \alpha \text{ iff } \Gamma \vdash_{\mathbf{Cila}} \alpha.$$

As a consequence of this, the following holds, for every $\Gamma \cup \{\alpha\} \subseteq For^{\circ}$:

$$\Gamma \vdash_{\mathbf{Cila}} \alpha \text{ implies } \Gamma^{\star} \vdash_{\mathbf{Cila}} \alpha^{\star}.$$

Proof. An easy extension of the proof of Theorem 106. In fact, taking into account that C_1 coincides with **Cila**$^{\copyright}$ (recall Remark 109) and also the fact that axioms (ca1)–(ca3) of Definition 108 are translated by $\star$ in terms of the homonymous axioms of Definition 28, the proof is immediate. ■

REMARK 111. Consider the logic **Cl** obtained from **Cil** by removing axiom (ci). In other words, **Cl** is defined by axiom schemas (Ax1)–(Ax11) (see Definition 28), (cl) (see Definition 105), plus (MP). Let **Cil**$^{\copyright}$ be the logic defined in Theorem 106. It is easy to check, though, that the results in Theorem 106 are still valid if we uniformly substitute **Cl** for **Cil**. The logic **Cla**, studied in [Avron, 2005b], may now be obtained from **Cl** by adding axiom schemas (ca1)–(ca3) of Definition 108, and the proof of Theorem 110 is still valid if if we uniformly substitute **Cla** for **Cila**.[9] However, according to Definition 33, we can say that **Cila** corresponds to C_1, but we cannot say the same about the **C**-system **Cla**. ■

Taking into account the new axioms from Definition 108, it is easy to prove in **Cia** the following particular version of a Derivability Adjustment

[9] We thank Arnon Avron for pointing this fact to us.

Theorem (recall Remark 26 and compare the following with what was said at the beginning of Subsection 3.6):

THEOREM 112.
Let Π denote the set of atomic formulas occurring in $\Gamma \cup \{\alpha\}$.
Then, $\Gamma \vdash_{\mathbf{CPL}} \alpha$ iff there is some $\Delta \subseteq \Pi$ such that $\circ(\Delta), \Gamma \vdash_{\mathbf{Cia}} \alpha$.

Proof. Recall that **CPL** may be axiomatized by (Ax1)–(Ax11), (MP) and the 'explosion law': (exp) $\alpha \rightarrow (\neg\alpha \rightarrow \beta)$. Consider some $\Gamma \cup \{\alpha\} \subseteq For$ such that $\Gamma \vdash_{\mathbf{CPL}} \alpha$. By induction on the length n of a derivation in **CPL** of α from Γ it will be proven that $\circ(\Delta), \Gamma \vdash_{\mathbf{Cia}} \alpha$ for some $\Delta \subseteq \Pi$. If $n = 1$ then either $\alpha \in \Gamma$ or α is an instance of an axiom of **CPL**. In the first case the proof is trivial. In the second case, there is just one case in which α is not an axiom of **Cia**, namely, when α is an instance $\delta \rightarrow (\neg\delta \rightarrow \beta)$ of (exp). Let Δ be the set of propositional variables occurring in δ. Then, by Theorem 101(i) and by (ca1)–(ca3), it is easy to prove (by induction on the complexity of δ) that $\circ(\Delta) \vdash_{\mathbf{Cia}} \circ\delta$. On the other hand, from (bc1) and (MP) it follows that $\circ\delta \vdash_{\mathbf{Cia}} \alpha$. Thus $\circ(\Delta), \Gamma \vdash_{\mathbf{Cia}} \alpha$, where $\Delta \subseteq \Pi$. Suppose now that α follows from β and $\beta \rightarrow \alpha$ by (MP), in the last step of a given derivation in **CPL** of α from Γ. By induction hypothesis, $\circ(\Delta_1), \Gamma \vdash_{\mathbf{Cia}} \beta$ and $\circ(\Delta_2), \Gamma \vdash_{\mathbf{Cia}} \beta \rightarrow \alpha$ for some $\Delta_1, \Delta_2 \subseteq \Pi$. Thus $\circ(\Delta_1), \circ(\Delta_2), \Gamma \vdash_{\mathbf{Cia}} \alpha$, by (MP). But of course $\circ(\Delta_1) \cup \circ(\Delta_2) = \circ(\Delta_1 \cup \Delta_2)$, so we have that $\circ(\Delta_1 \cup \Delta_2), \Gamma \vdash_{\mathbf{Cia}} \alpha$, and that concludes the first half of the proof.

Conversely, suppose now that $\Gamma \cup \{\alpha\} \subseteq For$ is such that $\circ(\Delta), \Gamma \vdash_{\mathbf{Cia}} \alpha$ for some $\Delta \subseteq \Pi$. If $\Gamma \nvdash_{\mathbf{CPL}} \alpha$ then there exists a classical valuation v such that $v(\Gamma) \subseteq \{1\}$ and $v(\alpha) = 0$. Extend v to For° by putting $v(\circ\beta) = 1$ for every $\beta \in For^\circ$. Then v is a model for **Cia** such that $v(\circ(\Delta) \cup \Gamma) \subseteq \{1\}$, therefore $v(\alpha) = 1$, a contradiction. Thus $\Gamma \vdash_{\mathbf{CPL}} \alpha$. ■

As pointed out already in [da Costa, 1963] and [da Costa, 1974], the same result holds good for any logic C_n, assuming in each case the appropriate definition of $\circ\alpha$.

Recalling that **eCPL** is just the classical propositional logic **CPL** plus the axiom schema $\circ\alpha$, we may also propose the following alternative way of recovering classical reasoning inside our present **LFI**s:

THEOREM 113. Consider the mapping t_3: $For^\circ \longrightarrow For^\circ$, recursively defined as follows:

1. $t_3(p) = \circ p$, for every $p \in \mathcal{P}$;
2. $t_3(\gamma \# \delta) = (t_3(\gamma) \# t_3(\delta))$, if $\# \in \{\wedge, \vee, \rightarrow\}$;
3. $t_3(\divideontimes\gamma) = \divideontimes t_3(\gamma)$, if $\divideontimes \in \{\neg, \circ\}$.

Then t_3 conservatively translates **eCPL** inside of **Cia**.

Proof. Using compactness, the deduction theorem, and the definition of t_3, it is enough to prove that $\vdash_{\mathbf{eCPL}} \alpha$ iff $\vdash_{\mathbf{Cia}} t_3(\alpha)$ for every α in For°.

We first prove from left to right. Given a formula $\alpha(p_1, \ldots, p_n)$ in For°, then $t_3(\alpha) = \alpha(\circ p_1, \ldots, \circ p_n)$. From this, using axioms (ca1)–(ca3), axiom $(\mathrm{cc})_n$ (Definition 75) and Theorem 101(i) it is not hard to prove by induction on the complexity $\ell(\alpha)$ of α that $\vdash_{\mathbf{Cia}} \circ t_3(\alpha)$ for every $\alpha \in For^\circ$. Observe that, if β is an axiom of **eCPL** different from (exp) (see Remark 29) then $t_3(\beta)$ is a theorem of **Cia**. On the other hand, if $\beta = \delta \to (\neg\delta \to \gamma)$ is an instance of (exp) then $t_3(\beta) = t_3(\delta) \to (\neg t_3(\delta) \to t_3(\gamma))$, and the latter is provable in **Cia** from (bc1) and $\circ t_3(\delta)$. Thus, $t_3(\beta)$ is a theorem of **Cia**. Note also that any application of *modus ponens* in **eCPL** is transformed into an application of *modus ponens* in **Cia**. Consequently, given a derivation $\alpha_1, \ldots, \alpha_n = \alpha$ of α in **eCPL**, the finite sequence of formulas $t_3(\alpha_1), \ldots, t_3(\alpha_n) = t_3(\alpha)$ may be transformed into a derivation of $t_3(\alpha)$ in **Cia**. This shows that $\vdash_{\mathbf{eCPL}} \alpha$ implies $\vdash_{\mathbf{Cia}} t_3(\alpha)$.

In order to prove the converse, consider the definition of an adequate bivaluation semantics for **Cia**, adding to the clauses of a bivaluation semantics for **Ci** (see Definition 84) the clause $(vC7)$ of Example 65. Now, given an **eCPL**-valuation v, consider the mapping $v' \colon \mathcal{P} \cup \{\neg p : p \in \mathcal{P}\} \longrightarrow \mathbf{2}$ such that $v'(p) = 1$ for every $p \in \mathcal{P}$, and $v'(\neg p) = 1$ iff $v(p) = 0$. Define now $v'(\circ p) = 1$ iff $v'(\neg p) = 0$, and extend v' homomorphically to the remaining formulas in For° using the truth-tables for **eCPL**. That is, for formulas other than p, $\neg p$ and $\circ p$ (for $p \in \mathcal{P}$) the mapping v' is defined as a classical valuation and moreover satisfies $v'(\circ\alpha) = 1$ for every non-atomic α. It is easy to see that this v' is indeed a **Cia**-valuation. An induction on the complexity $\ell(\alpha)$ of α shows that $v(\alpha) = v'(t_3(\alpha))$ for every $\alpha \in For^\circ$. Finally, suppose that $\nvdash_{\mathbf{eCPL}} \alpha$. Then, there is some **eCPL**-valuation v such that $v(\alpha) = 0$. But then, by the above argument, there is some **Cia**-valuation v' such that $v'(t_3(\alpha)) = 0$ and so $\nvdash_{\mathbf{Cia}} t_3(\alpha)$. ■

Straightforward adaptations of the above argument show that the same t_3 acts as a conservative translation between **eCPL** and all logics defined in item (ii) of Definition 108. So, in order to perform 'classical inferences' within such logics (and even within C_1, in view of Theorem 110), it suffices to translate every atomic formula p into $\circ p$.

Axioms (ca1)–(ca3) of Definition 108 describe a certain form of propagation of consistency through conjunction. There are several other sensible ways of allowing consistency or inconsistency to propagate. For instance, it also makes sense to think of propagation of consistency through disjunction:

DEFINITION 114.
(i) The logic **Cio** is obtained by the addition to **Ci** of the axiom schemas:

(co1) $(\circ\alpha \vee \circ\beta) \to \circ(\alpha \wedge \beta)$;

(co2) $(\circ\alpha \vee \circ\beta) \to \circ(\alpha \vee \beta)$;

(co3) $(\circ\alpha \vee \circ\beta) \rightarrow \circ(\alpha \rightarrow \beta)$.

(ii) The logic **Cilo** is obtained by the addition to **Cio** of the axiom schema (cl) or, equivalently, by the addition of axioms (co1)–(co3) to **Cil** (see Definition 105). ■

The logic $\mathbf{Cilo}^{\copyright}$, the version of **Cilo** over signature Σ (using $\mathbf{Cil}^{\copyright}$ instead of **Cil**, see Theorem 106), was introduced in [Béziau, 1990] and was studied under the name C_1^+ in [da Costa *et al.*, 1995]. As in Definition 108, several other logics may be defined extending **Cio** by tinkering with axioms (cf), (cb) and (ce).

Obviously, C_1^+ is a deductive extension of C_1. Its characteristic weaker requirement to obtain consistency of a complex formula, namely, the consistency of at least one of its components, reflects in the following immediate stronger result:

THEOREM 115.
If $\Gamma \vdash_{\mathbf{Cio}} \circ\beta$ for some subformula β of α, then $\Gamma \vdash_{\mathbf{Cio}} \circ\alpha$. ■

An argument similar to the one presented in the proof of Theorem 113 will show again that the same t_3 defines also a conservative translation between **eCPL** and the logics presented in Definition 114.

On what concerns the interdefinability of the binary connectives with the help of our primitive paraconsistent negation (compare with Theorem 64), one can now count on the following extra rules:

THEOREM 116.
In **Cia** the following holds good:
(ix) $\neg(\neg\alpha \wedge \neg\beta) \vdash_{\mathbf{Cia}} (\alpha \vee \beta)$.
In **Cio** the following hold good:
(vi) $\neg(\alpha \wedge \neg\beta) \vdash_{\mathbf{Cio}} (\alpha \rightarrow \beta)$;
(vii) $\neg(\alpha \rightarrow \beta) \vdash_{\mathbf{Cio}} (\alpha \wedge \neg\beta)$;
(xi) $\neg(\neg\alpha \vee \neg\beta) \vdash_{\mathbf{Cio}} (\alpha \wedge \beta)$. ■

From Theorem 116(vii) and Theorem 52(ii) we can conclude that the replacement property (RP) (recall Remark 51) does not hold for any extension of **Cio**. However, a restricted form of this property may be recovered, in this specific case:

REMARK 117. Say that a logic **L** *allows for replacement with respect to* $\approx$ when $p_1 \approx p_2$ is a formula depending on the variables p_1 and p_2 such that, for every formula $\varphi(p_0, \ldots, p_n)$ and formulas $\alpha_0, \ldots, \alpha_n$, $\beta_0, \ldots, \beta_n$:

(RRP) $(\Vdash_{\mathbf{L}} \alpha_0 \approx \beta_0)$ and $\ldots$ and $(\Vdash_{\mathbf{L}} \alpha_n \approx \beta_n)$ implies
$\Vdash_{\mathbf{L}} \varphi(\alpha_0, \ldots, \alpha_n) \approx \varphi(\beta_0, \ldots, \beta_n)$.

Any such formula, when it exists, will be called a *congruence* of **L**. Notice that, for our present logics, full replacement holds exactly when $\leftrightarrow$ is a congruence. ■

In the case of C_1 (and, not surprisingly, also of **Cia**), it has been shown in [Mortensen, 1980] that no congruence exists distinct from the 'trivial' one, namely, the identity between formulas. The situation is different though in the case of **Cio** and its deductive extensions:

THEOREM 118. A congruence in **Cio** can be defined by setting $\alpha \approx \beta \stackrel{\text{def}}{=} (\alpha \leftrightarrow \beta) \wedge (\circ\alpha \wedge \circ\beta)$.

Proof. A semantic proof for **Cilo** was offered in Theorem 3.21 of [da Costa *et al.*, 1995]. A similar argument, adapted for **Cio**, can be found in Fact 3.81 of [Carnielli and Marcos, 2002]. ■

On what concerns the semantic presentation of the above logics, the following Theorems 121 and 125 exhibit sufficient conditions for showing that several of the logics mentioned so far fail to be characterizable by finite-valued truth-tables.

The first widely applicable theorem on non-characterizability by finite-valued truth-tables proceeds as follows. Consider the signature Σ°. Recall from Definition 28 that α^1 denotes the formula $\neg(\alpha \wedge \neg\alpha)$ and α^{n+1} abbreviates the formula $\neg(\alpha^n \wedge \neg\alpha^n)$ for $n \geq 1$. Consider, additionally, $\alpha^0 \stackrel{\text{def}}{=} \alpha$ for every α in For°. Finally, set $\delta(m) \stackrel{\text{def}}{=} (\bigwedge_{0 \leq i < m} \delta^i) \rightarrow \delta^m$ for $\delta \in For^\circ$ and $m \geq 1$.

LEMMA 119. Any set $\mathcal{M}$ of n-valued truth-tables for which positive classical logic ($\mathbf{CPL}^+$) or some deductive extension thereof is sound must validate all formulas of the form $\delta(m)$, for $m > n$.

Proof. The case $n < 2$ is obvious, for then $\mathcal{M}$ must be an adequate set of truth-tables for the trivial logic. The other cases are easy consequences of the Pigeonhole Principle of finite combinatorics and of the cyclic character of the composition of finite functions. Indeed, if $\mathcal{M}$ is n-valued, for some finite n, the truth-table determined by a formula δ^n must be identical to the truth-table of at least one among the formulas $\delta^0, \ldots, \delta^{n-1}$. But in that case, using classical properties of conjunction and implication, it follows that $\delta(m)$, and consequently $\delta(m)$, is valid according to $\mathcal{M}$. ■

The above lemma can be found at [Avron, 2007b]. The next result comes from [Marcos, 2005f].

LEMMA 120. No formula of the form $\delta(m)$ is derivable in the logic **Ciae**.

Proof. Consider, for $n \in \mathbb{N}$, the following sets $\mathcal{M}_n$ of infinitary truth-tables that take the truth-values from the ordinal $\omega + 1 = \omega \cup \{\omega\}$, where ω (the set of natural numbers) is the only undesignated truth-value:

$$x \wedge y = \begin{cases} 0, & \text{if } x = n \text{ and } y = n+1 \\ \mathsf{max}(x, y), & \text{otherwise} \end{cases}$$

$$x \vee y = \mathsf{min}(x, y)$$

$$x \to y = \begin{cases} \omega, & \text{if } x \in \mathbb{N} \text{ and } y = \omega \\ y, & \text{if } x = \omega \text{ and } y \in \mathbb{N} \\ 0, & \text{if } x = \omega = y \\ \mathsf{max}(x, y), & \text{otherwise} \end{cases}$$

$$\neg x = \begin{cases} \omega, & \text{if } x = 0 \\ 0, & \text{if } x = \omega \\ x + 1, & \text{otherwise} \end{cases} \qquad \circ x = \begin{cases} 0, & \text{if } x \in \{0, \omega\} \\ \omega, & \text{otherwise} \end{cases}$$

It is clear, on the one hand, that **Ciae** is sound for each $\mathcal{M}_n$. On the other hand, $\mathcal{M}_{2m+1}$ falsifies the formula $\delta(m+1)$. Indeed, consider an atomic sentence p in the place of δ and consider a valuation v such that $v(p) = 1$. It follows then that $v(p^i) = 2i + 1$, for $0 \leq i \leq m$, yet $v(p^{m+1}) = \omega$. But in that case $v(\delta(m+1)) = ((2m+1) \to \omega) = \omega$. ■

THEOREM 121. No **LFI** lying in between $\mathbf{CPL}^+$ and **Ciae** is finite-valued.

Proof. Suppose that **L** is a logic defined over Σ° lying in between $\mathbf{CPL}^+$ and **Ciae** such that **L** has an adequate finite-valued truth-functional semantics with, say, m truth-values. By Lemma 119 the formula $\delta(m+1)$ is valid with respect to this semantics and so it is a theorem of **L**. But then $\delta(m+1)$ would be a theorem of **Ciae**, contradicting Lemma 120. ■

The previous result, albeit very general, does not cover cases of uncharacterizability by finite-valued truth-tables for logics satisfying the axiom (cl), for the truth-tables presented in Lemma 120 provide counter-models to this axiom. Here is, however, a similar argument that works fine in the latter case.

DEFINITION 122. Let $\mathbf{Cl}^-$ be the logic defined over the signature Σ° and obtained from **Cl** (see Remark 111) by removing axiom schemas (Ax10)–(Ax11). In other words, $\mathbf{Cl}^-$ is characterized by axiom schemas (Ax1)–(Ax9) (see Definition 28), (bc1) (see Definition 42), (cl) (see Definition 105), and the rule (MP). ■

Let δ_{ij}, for $i, j \neq 0$, denote the formula $\neg(p_i \wedge \neg p_j) \wedge (p_i \wedge \neg p_j)$, and let $\delta^{[n]}$ denote the disjunctive formula $\bigvee_{1 \leq i < j \leq n}(\delta_{ij} \to p_{n+1})$ for $n \geq 1$. Then:

LEMMA 123. Any set of n-valued truth-tables that is sound for the logic $\mathbf{Cl}^-$ must validate all formulas of the form $\delta^{[m]}$ for $m > n$.

Proof. Use the Pigeonhole Principle and the fact that

$$(\neg(\alpha \wedge \neg\alpha) \wedge (\alpha \wedge \neg\alpha)) \to \beta$$

may be derived from axioms (bc1), (cl) and the deduction theorem. ■

LEMMA 124. No formula of the form $\delta^{[n]}$ is derivable in the logic **Cilae**.

Proof. Use again the truth-tables in Lemma 120, but now simplify the table of conjunction as follows:

$$x \wedge y = \begin{cases} 0, & \text{if } y = x+1 \\ \mathsf{max}(x,y), & \text{otherwise} \end{cases}$$

It is routine to check that these truth-tables are sound for **Cilae**. Consider next a valuation v such that $v(p_i) = i$, for $i \leq n$, and $v(p_{n+1}) = \omega$. Then $v(\delta_{ij}) = j+2$ and so $v(\delta_{ij} \rightarrow p_{n+1}) = ((j+2) \rightarrow \omega) = \omega$ (for $1 \leq i < j \leq n$). Thus $v(\delta^{[n]}) = \omega$. ■

THEOREM 125. No **LFI** lying in between $\mathbf{Cl}^-$ and **Cilae** is finite-valued.

Proof. Analogous to the proof of Theorem 121, but now using formulas $\delta^{[n]}$, Lemma 123 and Lemma 124. ■

REMARK 126. A somewhat stronger version of Theorem 125 has recently been proven in [Avron, 2005b], where all logics in between $\mathbf{Cl}^-$ and **Cilae** are shown not to be characterizable even with the use of finite-valued non-deterministic truth-tables.

The logic **Cibae** (Definitions 108), an obvious extension of **Cila**, received an adequate interpretation in terms of possible-translations semantics in [Carnielli, 2000] and in [Marcos, 1999]. In the latter study, all the other logics from Definitions 108 and 114 have also received adequate possible-translations semantics. In [Avron, 2007a; Avron, 2005c; Avron, 2007b], even larger families of related logics have recently been given interpretations in terms of non-deterministic semantics, in a modular way. ■

We end this subsection with an axiomatization of two important 3-valued **LFI**s through the regulation of their ability to propagate inconsistency.

THEOREM 127. The logic **LFI1** described in Example 18 is axiomatized by adding to **Cie** (check Definition 100) the following axiom schemas:

(cj1) $\bullet(\alpha \wedge \beta) \leftrightarrow ((\bullet\alpha \wedge \beta) \vee (\bullet\beta \wedge \alpha))$

(cj2) $\bullet(\alpha \vee \beta) \leftrightarrow ((\bullet\alpha \wedge \neg\beta) \vee (\bullet\beta \wedge \neg\alpha))$

(cj3) $\bullet(\alpha \rightarrow \beta) \leftrightarrow (\alpha \wedge \bullet\beta)$

where, as usual, $\bullet\alpha$ is an abbreviation for $\neg\circ\alpha$. The logic $\mathbf{P}^1$ described in Example 19 is axiomatized by adding to **Ci** (check Definition 100) the following schema:

(cz) $\circ\alpha$ (for α non-atomic) ■

In the last theorem, note that (cz), in fact, consists of five axiom schemas, one for each connective in the signature Σ°, that is, (cz) is equivalent to the conjunction $\circ(\neg\alpha) \wedge \circ(\alpha \wedge \beta) \wedge \circ(\alpha \vee \beta) \wedge \circ(\alpha \rightarrow \beta) \wedge \circ(\circ\alpha)$. The logic $\mathbf{P}^1$ describes an extreme case of propagation of consistency into complex formulas, where no premises are needed so as to guarantee their consistency.

*5.3 **LFI**s that are maximal fragments of **CPL***

The paper [da Costa, 1974] suggested a list of 'natural' features that a paraconsistent logic should enjoy. One of these is that a paraconsistent logic should contain the most part of the schemas and rules of the classical propositional logic which do not interfere with paraconsistency. Following [Marcos, 2005d], one way of implementing this feature would be by requiring paraconsistent logics to be, in some specific sense, maximal deductive fragments of classical logic.

The following notion of maximality among logics may be used to analyze how close we are to having 'most of classical logic' inside paraconsistent systems:

DEFINITION 128. Let $\mathbf{L}1$ and $\mathbf{L}2$ be two logics written in the same signature. Then, $\mathbf{L}2$ is said to be *maximal relative to* $\mathbf{L}1$ if:
(i) $\mathbf{L}1$ is an extension of $\mathbf{L}2$;
(ii) if $\vdash_{\mathbf{L}1} \alpha$ but $\nvdash_{\mathbf{L}2} \alpha$, then the logic obtained from $\mathbf{L}2$ by adding α as a new axiom schema coincides with $\mathbf{L}1$.
When $\mathbf{L}1$ is clear from the context, we simply say that a logic $\mathbf{L}2$ satisfying conditions (i) and (ii) is *maximal*. ■

This notion of maximality is quite common in the literature.[10] It is well known, for instance, that each Łukasiewicz's logic $\mathrm{Ł}_m$, for $m > 2$, is maximal relative to **CPL** if and only if $(m-1)$ is a prime number. Also, **CPL** is maximal relative to the trivial logic, a logic in which all formulas are provable. On the other hand it is also well known that intuitionistic logic is not a maximal fragment of **CPL**, and there exists indeed an infinite number of intermediate logics between them. On what concerns the main **C**-systems presented this far, only the logic **LFI1** and the logic $\mathbf{P}^1$, described in Examples 18 and 19, and Theorem 127, are maximal relative to **CPL**, or relative to **eCPL**, the extended version of **CPL** introduced at the beginning of Subsection 3.6. In particular, the logic C_1 (or, equivalently, **Cila**$^{\copyright}$ — recall Remark 109), despite being the strongest logic introduced by da Costa on his first hierarchy of paraconsistent logics, is properly extended by $\mathbf{P}^1$

[10] Other notions of 'maximality' exist, such as the idea of defining maximal subsets of the classical entailment, considering not only valid formulas but valid inferences. That approach fails monotonicity, though, and the consequent 'maximal fragments' of classical logic do not define thus T-logics nor S-logics. We will make no development in the present paper in that direction, and choose rather to refer to the competent sources, such as [Batens, 1989] and [Batens, 1989].

and fails thus to be maximal with respect to classical logic. Therefore, none of the logics C_n presented in [da Costa, 1974] respects the requirement of containing the most part the schemas of classical logic, a requirement that may be found in that very same paper. Such an observation, in fact, is true also about the stronger logic called C_1^+ (or **Cilo**$^{\copyright}$), introduced after Definition 114.

Now we explore the intuitions underlying the 3-valued maximal **C**-systems $\mathbf{P}^1$ and **LFI1** showing how to generate a large class of related 3-valued maximal paraconsistent logics. Looking for models for contradictory and non-trivial theories, we start with non-trivial interpretations under which both some formula α and its negation $\neg\alpha$ would be simultaneously satisfied. A natural choice lies in the many-valued domain, more specifically in logics presented in terms of finite-valued truth-tables. Since we want to preserve classical theses as much as possible, the values of the connectives with classical (0 and 1) inputs will have classical outputs. Suppose we just introduce then an intermediate third value $\frac{1}{2}$, besides true (1) and false (0), fixing $D = \{1, \frac{1}{2}\}$ as the set of designated values. Then there are two possible classic-like truth-tables for a negation validating α and $\neg\alpha$ simultaneously, for some α, namely:

	$\neg$
1	0
$^1/_2$	$^1/_2$ or 1
0	1

With respect to the other connectives of the signature Σ (since we try to keep them as classical as possible), we add now the following higher-level classic-like requirements:

(C$\wedge$) $(x \wedge y) \in D$ iff $x \in D$ and $y \in D$;
(C$\vee$) $(x \vee y) \in D$ iff $x \in D$ or $y \in D$;
(C$\rightarrow$) $(x \rightarrow y) \in D$ iff $x \notin D$ or $y \in D$.

The above constraints leave us with the following options:

$\wedge$	1	$^1/_2$	0
1	1	$^1/_2$ or 1	0
$^1/_2$	$^1/_2$ or 1	$^1/_2$ or 1	0
0	0	0	0

$\vee$	1	$^1/_2$	0
1	1	$^1/_2$ or 1	1
$^1/_2$	$^1/_2$ or 1	$^1/_2$ or 1	$^1/_2$ or 1
0	1	$^1/_2$ or 1	0

$\rightarrow$	1	$^1/_2$	0
1	1	$^1/_2$ or 1	0
$^1/_2$	$^1/_2$ or 1	$^1/_2$ or 1	0
0	1	$^1/_2$ or 1	1

This yields 2^3 options for conjunctions, 2^5 options for disjunctions, 2^4 options for implications, and, as stated above, 2^1 options for negations, adding up to 2^{13} $(= 8,192)$ possible logics to deal with, in the signature Σ. Of course, not all those logics are necessarily 'interesting'. We can upgrade each of those logics into an **LFI** by considering the signature $\Sigma^{\circ\bullet}$ and adding the following tables for consistency and inconsistency operators:

	$\circ$	$\bullet$
1	1	0
$^1/_2$	0	1
0	1	0

This means that the consistent models are the ones characterized by classical valuations, and only those. Notice that, in the above truth-tables, $\circ$ can be defined by setting $\circ\alpha \stackrel{\text{def}}{=} \neg\bullet\alpha$ or, alternatively, $\bullet$ can be defined by setting $\bullet\alpha \stackrel{\text{def}}{=} \neg\circ\alpha$.

DEFINITION 129. Fix Σ as any one among the signatures Σ°, $\Sigma^\bullet$ or $\Sigma^{\circ\bullet}$. The collection of logics over Σ defined by the above truth-tables, with designated values $D = \{1, \frac{1}{2}\}$, will be called $8Kb$. Each logic in this collection makes up a choice as to which truth-table for negation, for conjunction, for disjunction and for implication it will adopt. ■

Clearly, every logic in $8Kb$ is a fragment of **eCPL**, the extended classical propositional logic, if we consider in **eCPL** the usual definition of the inconsistency connective as the negation of the consistency connective. Note also that the logic Pac (see Example 17) does not belong to $8Kb$, because it cannot define the connectives $\circ$ and $\bullet$. On the other hand, its conservative extension **LFI1** contains those connectives, and as a matter of fact the latter logic belongs to $8Kb$. The 3-valued logic $\mathbf{P}^1$ also belongs to $8Kb$, and we already know that these two logics are axiomatizable by the addition of suitable axioms to the axiomatization of **Ci** (see Theorem 127). As shown in [Marcos, 2000], this same method may be extended to the whole $8Kb$:

THEOREM 130. (i) Every logic in $8Kb$ is an axiomatic extension of **Cia**.
(ii) All the logics in $8Kb$ are distinct from each other, and they are all maximal relative to **eCPL**.
(iii) All the logics in $8Kb$, and their fragments, are boldly paraconsistent. ■

It is just a combinatorial divertissement to check the following facts:

THEOREM 131. All the $8,192$ logics in $8Kb$ are **C**-systems based on **CPL** and extending **Cia** (cf. Definition 108). Out of these, $7,680$ are in fact **dC**-systems, being able to define $\circ$ and $\bullet$ in terms of the other connectives (all being, therefore, maximal relative to **CPL**, and not only to **eCPL**). Of these, $4,096$ are able to define $\circ\alpha$ as $\neg(\alpha \wedge \neg\alpha)$, and so all of them

do extend C_1 (that is, $\mathbf{Cila}^{\copyright}$). Of the 7,680 logics which are **dC**-systems, 1,680 extend **Cio** (cf. Definition 114), and 980 of the latter are able to define $\circ\alpha$ as $\neg(\alpha \wedge \neg\alpha)$, and so all these 980 logics extend C_1^+ (that is, $\mathbf{Cilo}^{\copyright}$). ■

REMARK 132. The reader should bear in mind that, in view of Definition 27, if we want to prove that a given logic **L**2 is a **C**-system based on another logic **L**1, we might have to adjust its signature Σ_2 by adding definable connectives so as to guarantee that it will extend the signature Σ_1 of **L**1 (as it was done, for instance, in the proof of Theorem 44). In contrast to this, in view of Definition 128, if we want to prove that **L**2 is maximal relative to a logic **L**3, it might be necessary to adjust the signatures of both logics so that they coincide. Such signature adjustments are tacitly assumed in the statements of Theorems 130 and 131. So, in more practical terms, in order to prove that a given logic **L** in $8Kb$ is a **C**-system based on **CPL** we ought to add to its signature a new symbol for a (definable) classical negation. On the other hand, in order to prove that **L** is maximal relative to classical logic we had better assume in general that the latter logic is presented as **eCPL**, using the signature Σ° of Remark 15. In case **L** is a **dC**-system, then it will suffice to consider classical logic presented as **CPL**, and write **L** in the signature Σ, letting $\circ$ and $\bullet$ be introduced, in each case, by their circumstantial definitions. ■

The replacement property (RP) had already been shown to fail for our foremost logic samples from the $8Kb$. Indeed, the proof of items (iv) and (v) of Theorem 50 showed that both **LFI1** and $\mathbf{P}^1$ fail (RP). This negative feature may be generalized, as shown in [Marcos, 2000]:

THEOREM 133. (RP) cannot hold in any of the logics in $8Kb$.

Proof. This is true in general for any extension of **Cia**, as we may conclude from Theorem 81(ii) and Theorem 116(ix). To complete the proof, recall Theorem 130(i).

You will also be able to check the above result, alternatively, using the classical negation below, whose truth-table could already be found in Example 17 (check also Theorem 134), together with the result in Theorem 52(a)(i). ■

As a consequence of Theorem 133 the logics in $8Kb$ are not suitable to an algebraization by means of a direct Lindenbaum-Tarski-style procedure. However, the following results guarantee that all of them are algebraizable in the sense of Blok-Pigozzi (cf. [Blok and Pigozzi, 1989]).

THEOREM 134. Each one of the logics in $8Kb$ defines the following truth-table for classical negation and at least one of the two congruences below:

	$\sim$
1	0
$^1/_2$	0
0	1

$\equiv$	1	$^1/_2$	0
1	1	0	0
$^1/_2$	0	$^1/_2$ or 1	0
0	0	0	1

Proof. It is possible to define $\bot$ either as $(\alpha \wedge (\neg\alpha \wedge \circ\alpha))$ or as $(\circ\alpha \wedge \neg\circ\alpha)$, for any formula α. Then, we can define $\sim\alpha$ either as $(\neg\alpha \wedge \circ\alpha)$ or as $(\alpha \rightarrow \bot)$. One of the above congruences $(\alpha \equiv \beta)$ can always be defined by $((\alpha \leftrightarrow \beta) \wedge (\circ\alpha \leftrightarrow \circ\beta))$. In case we prefer to have $(\frac{1}{2} \equiv \frac{1}{2}) = 1$, we can assure that we define this specific congruence by setting $(\alpha \bowtie \beta) \stackrel{\text{def}}{=} \sim\sim(\alpha \equiv \beta)$. ■

The following theorem generalizes a result obtained in [Lewin *et al.*, 1990] for the logic $\mathbf{P}^1$:

THEOREM 135. All the logics in $8Kb$ are Blok-Pigozzi algebraizable.

Proof. Consider $\Delta(p_0, p_1) = \{(p_0 \equiv p_1)\}$ or $\Delta = \{(p_0 \bowtie p_1)\}$, where $\equiv$ and $\bowtie$ are defined as in the proof of the Theorem 134. Consider the sets

$$\delta(p_0) = \{((p_0 \rightarrow p_0) \rightarrow p_0)\}, \quad \varepsilon(p_0) = \{(p_0 \rightarrow p_0)\}$$

and check that the corresponding algebraizability conditions of [Blok and Pigozzi, 1989] are satisfied. ■

On what concerns the expressibility spectrum of the class $8Kb$ and of the distinguished logics $\mathbf{P}^1$ and **LFI1**, the following results can be checked:

THEOREM 136.
(i) The truth-tables of $\mathbf{P}^1$ can be defined inside of any of the logics in $8Kb$.
(ii) All the truth-tables in $8Kb$ can be defined inside of **LFI1**.

Proof. Item (i). Fix some logic $\mathbf{L}$ belonging to $8Kb$. Let $\wedge, \vee, \rightarrow, \neg, \circ$ and $\bullet$ be its primitive connectives, and let $\sim$ be the classical negation defined inside $\mathbf{L}$ as in Theorem 134. Then, the $\mathbf{P}^1$-negation of a formula α may be defined in $\mathbf{L}$ as $\sim\sim\neg\alpha$. The $\mathbf{P}^1$-conjunction of some given formulas α and β may be defined in $\mathbf{L}$ either as $\sim\sim(\alpha \wedge \beta)$ or as $(\sim\sim\alpha \wedge \sim\sim\beta)$. A definition in the same vein applies to both disjunction and implication. Note that the truth-tables in $\mathbf{L}$ for the connectives $\circ$ and $\bullet$ already coincide with those of $\mathbf{P}^1$.
Item (ii). A proof of this property may be found in [Avron, 1999]. A constructive proof may be found in [Marcos, 1999] and [Carnielli *et al.*, 2000]. ■

COROLLARY 137. (i) The logic $\mathbf{P}^1$ can be conservatively translated into any of the logics in $8Kb$. (ii) Any of the logics in $8Kb$ can be conservatively translated into **LFI1**. ■

As argued in [Avron, 1991], the logic **LFI1** has several properties that justify its role as one of the most 'natural' 3-valued paraconsistent logics. Theorem 136(ii) and Corollary 137(ii) show already how linguistically and deductively expressive this logic is.

A last note on algebraization. We had the chance in several occasions above to witness how replacement fails for many of our **LFI**s. This often makes it difficult to provide algebraic counterparts, in the usual sense, for those logics. However, it is interesting to observe that a kind of algebraic treatment for some wilder **C**-systems has been proposed and studied, for instance, in [Carnielli and de Alcantara, 1984] and [Seoane and de Alcantara, 1991] (for a partial survey, check the section 3.12 of [Carnielli and Marcos, 2002]). Additionally, an approach for algebraizing **LFI**s based on an idea similar to that of a possible-translations structure was presented in [Bueno-Soler *et al.*, 2004] and [Bueno-Soler and Carnielli, 2005].

6 CONCLUSIONS AND FURTHER PERSPECTIVES

In this final part of this chapter we recall some definitions and results obtained and described above, and point to some interesting new problems and research directions connected to what has been presented.

From Section 3 on, some of the possibilities for the formalization and understanding of the relationship between the concepts of consistency, inconsistency, contradictoriness and triviality were explored at a very general and abstract level. Assuming that consistency could be expressed inside some paraconsistent logics, and assuming furthermore that the consistency of a given formula would legitimate its explosive character (that is, assuming (9), a so-called Gentle Principle of Explosion), we have presented in Subsection 3.1 a general definition of a Logic of Formal Inconsistency, **LFI** (Definition 23). To actualize that definition (in a finitary way), we have started our study from the logic **mbC**, a very weak **C**-system based on classical logic (recall Definition 42), constructing all the remaining **C**-systems as extensions of **mbC**. Some specific extensions of **mbC** illustrated a subclass of the **C**-systems in which the connectives '$\circ$' for consistency and '$\bullet$' for inconsistency are expressible by means of other connectives. The members of this class were called **dC**-systems (recall Definition 32).

We briefly recall some consequences of our approach to formal (in)consistency: There are consistent and inconsistent logics. The inconsistent ones may be either paraconsistent or trivial, but not both. Let us say that a theory *has non-trivial models* only if these models do not assign designated values to all formulas. Thus, the theories of a consistent logic have non-trivial

models if and only if they are non-contradictory. Paraconsistent logics will tipically have non-trivial models for some of their contradictory theories. Paraconsistent logics may even have some trivial models among those models that satisfy contradictions. Such trivial models, however, cannot exist if the paraconsistent logics we are talking about are gently explosive, that is, if they constitute Logics of Formal Inconsistency. For each formula α of a logic $\mathbf{L}$, the consistency $\circ\alpha$ of α consists in the information that should be added to an α-contradictory theory in order to make it explosive, and consequently trivial. If the answer is 'nothing needs to be added', then α is already consistent in $\mathbf{L}$. This implies that, as expected, a logic is consistent if all of its formulas may be asserted to be consistent.

It will be clear now to the reader that there are many more examples of **C**-systems besides the logics C_n of da Costa and other logics axiomatized in a more or less similar fashion. The general idea is to express consistency and inconsistency inside a logic, at its object-language level. This approach allows us to collect in a single class of **LFI**s logics as diverse as the C_n, $\mathbf{P}^1$, $\mathbf{J}_3$ (renamed **LFI1**), and Jaśkowski's 'discussive' paraconsistent logic **D2** (cf. Example 24). Even normal modal logics in a convenient signature can be very naturally regarded as **dC**-systems. This bears on the relationship between negations and modalities, which reflects upon the possibilities of defining paraconsistent negations in modal environments, as studied by [Vakarelov, 1989], [Došen, 1986], [Béziau, 2002b], [Marcos, 2005e] and [Marcos, 2005b].

The fact that so many logics with diverse motivations and technical features may be recast as a **dC**-systems paves the way for an interesting question: To check whether other logics in the literature on paraconsistent logics could be characterized as **C**-systems, or, in general, as **LFI**s. Another related question is the following: How to enrich a given paraconsistent logic in order to turn it into an **LFI**? This was done by the logic **LFI1** (also known as **CLuNs**, or $\mathbf{J}_3$) with respect to the logic *Pac* (see Example 18). Consider now the 3-valued *closed set logic* studied in [Mortensen, 1995]. This logic consists of **LFI1**'s truth-tables of conjunction and of disjunction, plus the truth-table of negation of $\mathbf{P}^1$, where 0 is the only non-designated value. A consistency connective $\circ$ can then be defined via $\circ\alpha \stackrel{\text{def}}{=} \neg\neg(\alpha \vee \neg\alpha)$. The addition of an appropriate truth-table for implication would enrich the closed-set logic, and the resulting system would most certainly belong to the collection $8Kb$ of 3-valued maximal paraconsistent logics (recall Definition 129). But in that case, what would be the topological or set-theoretical significance of these new connectives?

The question of the duality between intuitionistic-like and paraconsistent logics, not explored in this chapter, is also worth mentioning. The concept of dual-intuitionism was already seized in the 40s by K. Popper, cf. [Popper, 1948], more or less at the same time as paraconsistency was being engendered. More recently, dual-intuitionism and dual-paraconsistency have

been studied, for example, in [Sylvan, 1990], [Urbas, 1996] and [Brunner and Carnielli, 2005]. The logics that are dual to paraconsistent are sometimes called 'paracomplete' (cf. [Loparić and da Costa, 1984]). Exploring the issue of duality, a natural question that appears concerns the notions that are dual to consistency and inconsistency, notions that one might dub 'determinedness' and 'undeterminedness'. Some initial explorations in that direction, and the related *Logics of Formal Undeterminedness*, may be found in [Marcos, 2005e].

Apparently, in the 40s, defenders of dual-intuitionism and paraconsistency independently realized that there should be a logic for general reasoning from hypotheses, accepting in certain cases some propositions and their negations as true (in the case of paraconsistency), or retaining some propositions and their negations as unfalsified (in the case of falsificationism). Indeed, there seems to be some common grounds connecting paraconsistency and the falsificationist program in Philosophy of Science, and that line of research seems worth pursuing. Similarly, paracomplete logics could have a contribution to make for the study of verificationism in science. The logical approach to such questions has recently been vindicated by studies such as [Shramko, 2005].

Applications of **LFI**s to yet other fields in philosophy seem promising. In [Costa-Leite, 2003] some possibilities of employing the connectives of consistency and inconsistency for the understanding of (and new regards on) epistemological problems related to the paradox of knowability are investigated. In [Marcos, 2005a] the use of a consistency-like modal connective for the modelling of the metaphysical notion of essence is tackled, and in that environment inconsistency turns out to mean a mere sort of 'accident'.

Another important issue concerns the incompleteness results in Arithmetic. Recall that Gödel's incompleteness theorems are based on the identification of 'consistency' and 'non-contradictoriness'. What would be the consequences if we started instead from the general notion of consistency hereby proposed (recall Definition 4)? Would it still be possible to reproduce Gödel's arguments? Quite possibly, his arguments would be rescued at the cost of assuming consistency (in our sense) of several formulas representing assumptions that would then become more explicit, and consequently open to debate. In the same spirit, it should be interesting to analyze the combination of **LFI**s with Modal Logics of Provability. In [Boolos, 1996], consistency is intended as a kind of opposite to the notion of provability. Using this idea, if the negation of a formula cannot be proven, then it is consistent with whatever else might be proven; a still weaker notion, connected to 'logical independence', would be to consider a formula to be consistent when neither this formula nor its negation can be proven. The insinuated exchange between Logics of Formal Inconsistency and Logics of Provability, in fact, seems attractive and deserves further research.

As it has been noted in the literature, it seems that most interesting prob-

lems related to paraconsistency appear already at the propositional level. It is possible though to extend a given propositional paraconsistent logic to higher orders using combination techniques such as fibring, if only we choose the right abstraction level to express our logics. See, for instance, [Caleiro and Marcos, 2001], where the logic C_1 is given a first-order version which coincides with the original one from [da Costa, 1963]. Another interesting possibility that involves first-order versions of paraconsistent logics in general, and especially of first-order **LFI**s, is the investigation of consistent yet ω-inconsistent theories (also related to Gödel's theorems).

Some other items for future research, already hinted at along the present text, are the following. From Theorem 79 we know that, in extensions of **mCi**, the formulas causing controllable explosion (Definition 9(ii)) coincide with the provably consistent formulas, that is, theorems of the form $\circ\alpha$. On the other hand, **mbC** does not have provably consistent formulas (see Theorem 47). So, is the logic **mbC** (see Definition 42) *not* controllably explosive? On another trail, we have seen that there are extensions of **mbC** for which the replacement property holds good (see Remark 53), and we have seen that to find extensions of **mCi** with that same property all one needs to do is to devise logics that respect a certain rule (EC) (see Subsection 3.2 and Theorem 82). Can we circumvent negative results such as Theorems 52 and 81 and find interesting extensions of **mCi** enjoying the replacement property (RP)? At any rate, turning the attention to extensions of **mbC** that do not extend **mCi** but that do enjoy (RP) is a feasible enterprise (recall Remark 53), and it seems indeed to be a very attractive one, still to be further developed. On yet another direction, what other uses could we give to our semantical tools (valuations and possible-translations semantics)? The results about uncharacterizability by finite-valued truth-tables in Theorems 121 and 125 are very powerful and widely applicable, but they cannot help us in proving that logics such as **Cioe** do not have adequate finite-valued truth-tables. Can we find other flexible and wide-ranging similar results to the same effect?[11]

Finally, we have started our work in this chapter from a traditional abstract perspective. We have soon though shown that alternative semantical and proof-theoretical approaches were possible. In particular, we have given a few illustrations of a general method that permits us to deal with **C**-systems in terms of tableaux. The first wide-ranging method to such an effect was sketched in [Carnielli and Marcos, 2001b]. A more general method to obtain tableau procedures for logics endowed with a certain type of two-valued (even non-truth-functional) semantics was introduced in [Caleiro *et al.*, 2005b]. These techniques have been used here in Subsections 3.5 and 4.2 so as to obtain new adequate tableau systems for the logic C_1, as well as for

[11] It came to our notice that the problem concerning **Cioe** has recently been solved in [Avron, 2007b], where in fact all logics in between $\mathbf{Cl}^-$ and **Ciboe** are shown not to be characterizable with the use of finite-valued non-deterministic truth-tables.

mbC and **mCi**. The possibility of further exploring and refining this kind of approach seems promising for applications of **LFI**s in database theory (see Example 18), an area of research critically sensible to the presence of contradictions.

7 LIST OF AXIOMS AND SYSTEMS

We list here all the main principles, axioms and systems studied throughout the chapter, indicating the place where they were introduced in the text.

Principles

(1) Principle of Non-Contradiction : Subsection 2.1
(2) Principle of Non-Triviality : Subsection 2.1
(3) Principle of Explosion, or *Pseudo-Scotus*, or *Ex Contradictione Sequitur Quodlibet* : Subsection 2.1
(4) Paraconsistent logic (first definition) : Subsection 2.2
(5) Paraconsistent logic (second definition) : Subsection 2.2
(6) Paraconsistent logic (third definition) : Subsection 2.2
(7) Principle of *Ex Falso Sequitur Quodlibet* : Subsection 2.2
(8) Supplementing Principle of Explosion : Subsection 2.2
(9) Gentle Principle of Explosion : Subsection 3.1
(10) Finite Gentle Principle of Explosion : Subsection 3.1

Axioms, Rules and Metaproperties

(Ax1)–(Ax11) : Definition 28
(bc1) : Definition 28, Definition 42
$(bc1)'$: Subsection 4.3
$(bc1)''$: Theorem 106
(ca1)–(ca3) : Definition 28, Definition 108
(cb) : Subsection 5.1
$(cc)_n$: Definition 75
$(cc)'_n$: Subsection 4.3
(cd) : Subsection 5.1
(ce) : Subsection 4.4
(cf) (= (Ax11)) : Subsection 4.4
(ci) : Definition 75
$(ci)'$: Subsection 4.3
(cj1)–(cj3) : Theorem 127
(cl) : Definition 105
(co1)–(co3): Definition 114
(Con1)–(Con6) : Subsection 2.1
(cz) : Theorem 127
(EC) : Subsection 3.2

Systems

LFI1 : Example 18, Theorem 127
$\mathcal{M}_0$: Subsection 3.4
$\mathcal{M}_1$: Subsection 4.2
mbC : Definition 42
mbCe : Definition 100
mCi : Definition 75
mCi$^\bullet$: Definition 97
mCi$^{\circ\bullet}$: Subsection 4.3
mCie : Definition 100
MIL : Example 10
$\mathbf{P}^1$: Example 19, Theorem 127
Pac : Example 17
PI : Definition 36

8 ACKNOWLEDGEMENTS

The first and second authors acknowledge support by CNPq, Brazil, through individual research grants. The third author was supported by the Fundação para a Ciência e a Tecnologia (Portugal) and FEDER (European Union), via the grant SFRH / BD / 8825 / 2002 and the CLC / IST. This research was also supported by *Fundação de Amparo à Pesquisa do Estado de São Paulo* (FAPESP), Brazil, Thematic Project number 2004/14107-2 ("ConsRel"), and partially supported by FCT and UE FEDER POCI via SQIG at IT (Portugal). The authors are grateful to all colleagues who have helped them to clarify this material on several occasions, and specially to Arnon Avron, Manuel Bremer, Graham Priest, Casey McGinnis and Carlos Caleiro for detailed criticism and feedback that led to substantial improvement in this chapter. It goes without saying that their help does not necessarily mean that they endorse our views, nor are they responsible for any mistakes or imprecisions that might still lurk here unfound.

BIBLIOGRAPHY

[Arruda, 1975] A. I. Arruda. Remarques sur les systèmes C_n. *Comptes Rendus de l'Académie de Sciences de Paris (A-B)*, 280:1253–1256, 1975.

[Arruda, 1980] A. I. Arruda. A survey of paraconsistent logic. In A. I. Arruda, R. Chuaqui, and N. C. A. da Costa, editors, *Mathematical Logic in Latin America: Proceedings of the* IV *Latin American Symposium on Mathematical Logic*, pages 1–41. North-Holland, 1980.

[Asenjo, 1966] F. G. Asenjo. A calculus of antinomies. *Notre Dame Journal of Formal Logic*, 7:103–105, 1966.

[Avron, 1986] A. Avron. On an implication connective of *RM*. *Notre Dame Journal of Formal Logic*, 27:201–209, 1986.

[Avron, 1991] A. Avron. Natural 3-valued logics — Characterization and proof theory. *The Journal of Symbolic Logic*, 56(1):276–294, 1991.

[Avron, 1999] A. Avron. On the expressive power of three-valued and four-valued languages. *Journal of Logic and Computation*, 9:977–994, 1999.

[Avron, 2002] A. Avron. On negation, completeness and consistency. In D. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic, 2nd Edition*, volume 9, pages 287–319. Kluwer Academic Publishers, 2002.

[Avron, 2005a] A. Avron. Non-deterministic matrices and modular semantics of rules. In J.-Y. Béziau, editor, *Logica Universalis*, pages 149–167. Birkhäuser Verlag, Basel, Switzerland, 2005.

[Avron, 2005b] A. Avron. Non-deterministic semantics for paraconsistent **C**-systems. In L. Godo, editor, *Proceedings of the* VIII *European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty* (ECSQARU 2005), held in Barcelona, ES, July 2005, volume 3571 of *Lecture Notes in Computer Science*, pages 625–637. Springer, 2005.

[Avron, 2005c] A. Avron. Logical non-determinism as a tool for logical modularity: An introduction. In S. N. Artemov, H. Barringer, A. S. d'Avila Garcez, L. C. Lamb, J. Woods, editors, *We Will Show Them! Essays in honour of Dov Gabbay*, volume 1, pages 105–124. College Publications, 2005.

[Avron, 2007a] A. Avron. Non-deterministic semantics for families of paraconsistent logics In Béziau et al. [2007].

[Avron, 2007b] A. Avron. Non-deterministic semantics for logics with a consistency operator. Forthcoming.

[Avron and Lev, 2005] A. Avron and I. Lev. Non-deterministic multiple-valued structures. *Journal of Logic and Computation*, 15:241–261, 2005.

[Batens, 1980] D. Batens. Paraconsistent extensional propositional logics. *Logique et Analyse*, 90/91:195–234, 1980.

[Batens, 1989] Diderik Batens. Dynamic dialectical logics. In Priest et al. [1989], pages 187–217.

[Batens, 1989] Diderik Batens. *Logica Universalis*, 1:221–242, 2007.

[Batens and De Clercq, 2000] D. Batens and K. De Clercq. A rich paraconsistent extension of full positive logic. *Logique et Analyse (N.S.)*, 185/188:227–257, 2004.

[Béziau, 1990] J.-Y. Béziau. Logiques construites suivant les méthodes de da Costa. I. Logiques paraconsistantes, paracompletes, non-alèthiques construites suivant la première méthode de da Costa. *Logique et Analyse (N.S.)*, 131/132:259–272, 1990.

[Béziau, 1994] J.-Y. Béziau. Théorie legislative de la négation pure. *Logique et Analyse*, 147/148:209–225, 1994.

[Béziau, 1999] J.-Y. Béziau. La véritable portée du théorème de Lindenbaum-Asser. *Logique et Analyse*, 167/168:341–359, 1999.

[Béziau, 2002a] J.-Y. Béziau. Are paraconsistent negations negations? In W. A. Carnielli, M. E. Coniglio, and I. M. L. D'Ottaviano, editors, *Paraconsistency - the Logical Way to the Inconsistent*, volume 228 of *Lecture Notes in Pure and Applied Mathematics*, pages 465–486, New York, 2002. Marcel Dekker.

[Béziau, 2002b] J.-Y. Béziau. $S5$ is a paraconsistent logic and so is first-order classical logic. *Logical Studies*, 9:301–309, 2002.

[Béziau *et al.*, 2007] J.-Y. Béziau, W. A. Carnielli and D. Gabbay, editors. *Handbook of Paraconsistency*, Proceedings of the III World Congress on Paraconsistency, held in Toulouse, FR, July 28–July 31, 2003, volume 4 of *Studies in Logic and Practical Reasoning*. Amsterdam: North-Holland, 2007 (in print).

[Blok and Pigozzi, 1989] W. J. Blok and D. Pigozzi. Algebraizable Logics. *Memoirs of the American Mathematical Society*, 396, 1989.

[Bobenrieth-Miserda, 1996] A. Bobenrieth-Miserda. *Inconsistencias ¿Por qué no? Un estudio filosófico sobre la lógica paraconsistente*. Tercer Mundo Editores, Santafé de Bogotá, 1996.

[Boolos, 1996] G. Boolos. *The Logic of Provability*. Cambridge University Press, 1996.

[Brunner and Carnielli, 2005] A. B. M. Brunner and W. A. Carnielli. Anti-intuitionism and paraconsistency. *Journal of Applied Logics*, 3(1):161–184, 2005.

[Bueno, 1999] O. Bueno. Truth, quasi-truth and paraconsistency. In W. A. Carnielli and I. M. L. D'Ottaviano, editors, *Advances in Contemporary Logic and Computer Science*, volume 235 of *Contemporary Mathematics Series*, pages 275–293. American Mathematical Society, 1999.

[Bueno-Soler *et al.*, 2004] J. Bueno-Soler, M. E. Coniglio, and W. A. Carnielli. Finite algebraizability via possible-translations semantics. In W. A. Carnielli, F. M. Dionísio, and P. Mateus, editors, *Proceedings of the Workshop on Combination of Logics: Theory and applications* (CombLog'04), held in Lisbon, PT, 28–30 July 2004, pages 79–86. Departamento de Matemática, Instituto Superior Técnico, 2004. Preprint available at URL = `http://www.cs.math.ist.utl.pt/comblog04/abstracts/bueno.pdf`.

[Bueno-Soler and Carnielli, 2005] J. Bueno-Soler and W. A. Carnielli. Possible-translations algebraization for paraconsistent logics. *Bulletin of the Section of Logic*, 34(2):77–92, 2005. Preprint available at *CLE e-Prints*, vol. 5, n.6, 2005. URL = `http://www.cle.unicamp.br/e-prints/vol_5,n_6,2005.html`.

[Caleiro and Marcos, 2001] C. Caleiro and J. Marcos. Non-truth-functional fibred semantics. In H. R. Arabnia, editor, *Proceedings of the* 2001 *International Conference on Artificial Intelligence* (IC-AI'2001), volume II, pages 841–847. CSREA Press, 2001. Preprint available at URL = `http://wslc.math.ist.utl.pt/ftp/pub/CaleiroC/01-CM-fiblog10.ps`.

[Caleiro *et al.*, 2005a] C. Caleiro, W. Carnielli, M. E. Coniglio, and J. Marcos. Two's company: "The humbug of many logical values". In J.-Y. Béziau, editor, *Logica Universalis*, pages 169–189. Birkhäuser Verlag, Basel, Switzerland, 2005. Preprint available at URL = `http://wslc.math.ist.utl.pt/ftp/pub/CaleiroC/05-CCCM-dyadic.pdf`.

[Caleiro *et al.*, 2005b] C. Caleiro, W. A. Carnielli, M. E. Coniglio, and J. Marcos. How many logical values are there? Dyadic semantics for many-valued logics. Draft, 2005. Forthcoming.

[Caleiro *et al.*, 2005] C. Caleiro, W. Carnielli, J. Rasga, and C. Sernadas. Fibring of Logics as a Universal Construction. In D. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic, 2^{nd} Edition*, volume 13, pages 123–187. Springer, 2005.

[Carnielli, 1990] W. A. Carnielli. Many-valued logics and plausible reasoning. In *Proceedings of the* XX *International Symposium on Multiple-Valued Logic*, pages 328–335, Charlotte / NC, USA, 1990. IEEE Computer Society.

[Carnielli, 2000] W. A. Carnielli. Possible-translations semantics for paraconsistent logics. In D. Batens, C. Mortensen, G. Priest, and J. P. Van Bendegem, editors, *Frontiers of Paraconsistent Logic: Proceedings of the* I *World Congress on Paraconsistency*, Logic and Computation Series, pages 149–163. Baldock: Research Studies Press, King's College Publications, 2000.

[Carnielli and Coniglio, 2005] W. A. Carnielli and M. E. Coniglio. Splitting Logics. In S. Artemov, H. Barringer, A. Garcez, L. Lamb and J. Woods, editors, *We Will Show Them! Essays in Honour of Dov Gabbay*, volume 1, pages 389-414. College Publications, 2005.

[Carnielli and de Alcantara, 1984] W. A. Carnielli and L. P. de Alcantara. Paraconsistent algebras. *Studia Logica*, 43(1/2):79–88, 1984.

[Carnielli and Lima-Marques, 1992] W. A. Carnielli and M. Lima-Marques. Reasoning under inconsistent knowledge. *Journal of Applied Non-Classical Logics*, 2(1):49–79, 1992.

[Carnielli and Marcos, 1999] W. A. Carnielli and J. Marcos. Limits for paraconsistent calculi. *Notre Dame Journal of Formal Logic*, 40(3):375–390, 1999.

[Carnielli and Marcos, 2001a] W. A. Carnielli and J. Marcos. *Ex contradictione non sequitur quodlibet*. In *Proceedings of the Advanced Reasoning Forum Conference*, volume I of *Bulletin of Advanced Reasoning and Knowledge*, pages 89–109, 2001.

[Carnielli and Marcos, 2001b] W. A. Carnielli and J. Marcos. Tableau systems for logics of formal inconsistency. In H. R. Arabnia, editor, *Proceedings of the* 2001 *International Conference on Artificial Intelligence* (IC-AI'2001), volume II, pages 848–852. CSREA Press, 2001. URL = `http://tinyurl.com/7f2bh`.

[Carnielli and Marcos, 2002] W. A. Carnielli and J. Marcos. A taxonomy of **C**-systems. In W. A. Carnielli, M. E. Coniglio, and I. M. L. D'Ottaviano, editors, *Paraconsistency — The logical way to the inconsistent*, volume 228 of *Lecture Notes in Pure and Applied Mathematics*, pages 1–94, New York, 2002. Marcel Dekker. Preprint available at *CLE e-Prints*, 1(5), 2001.
URL = `http://www.cle.unicamp.br/e-prints/abstract_5.htm`.

[Carnielli *et al.*, 2000] W. A. Carnielli, J. Marcos, and S. de Amo. Formal inconsistency and evolutionary databases. *Logic and Logical Philosophy*, 8:115–152, 2000. Preprint available at *CLE e-Prints*, 1(6), 2001.
URL = `http://www.cle.unicamp.br/e-prints/abstract_6.htm`.

[Coniglio, 2005] M. E. Coniglio. Towards a stronger notion of translation between logics. *Manuscrito*, 28(2):231–262, 2005.

[Coniglio and Carnielli, 2002] M. E. Coniglio and W. A. Carnielli. Transfers between logics and their applications. *Studia Logica*, 72(3):367–400, 2002. Preprint available at *CLE e-Prints*, 1(4), 2001.
URL = `http://www.cle.unicamp.br/e-prints/abstract_4.htm`.

[Costa-Leite, 2003] A. Costa-Leite. Paraconsistency, modalities and cognoscibility (in Portuguese). Master's thesis, State University of Campinas (UNICAMP), Campinas, 2003.

[da Costa, 1959] N. C. A. da Costa. Observações sobre o conceito de existência em matemática. *Anuário da Sociedade Paranaense de Matemática*, 2:16–19, 1959.

[da Costa, 1963] N. C. A. da Costa. *Inconsistent Formal Systems* (in Portuguese), Habilitation Thesis, 1963. Republished by Editora UFPR, Curitiba, 1993.
URL = `http://www.cfh.ufsc.br/~nel/historia_logica/sistemas_formais.htm`.

[da Costa, 1974] N. C. A. da Costa. On the theory of inconsistent formal systems. *Notre Dame Journal of Formal Logic*, 15(4):497–510, 1974.

[da Costa and Alves, 1977] N. C. A. da Costa and E. Alves. A semantical analysis of the calculi C_n. *Notre Dame Journal of Formal Logic*, 18(4):621–630, 1977.

[da Costa and Béziau, 1993] N. C. A. da Costa and J.-Y. Béziau. Carnot's logic. *Bulletin of the Section of Logic*, 22(3):98–105, 1993.

[da Costa and Marconi, 1989] N. C. A. da Costa and D. Marconi. An overview of paraconsistent logic in the 80s. *The Journal of Non-Classical Logics*, 6(1):5–32, 1989.

[da Costa *et al.*, 1995] N. C. A. da Costa, J.-Y. Béziau, and O. A. S. Bueno. Aspects of paraconsistent logic. *Bulletin of the IGPL*, 3(4):597–614, 1995.

[da Silva *et al.*, 1999] J. J. da Silva, I. M. L. D'Ottaviano, and A. M. Sette. Translations between logics. In X. Caicedo and C. H. Montenegro, editors, *Models, Algebras and Proofs*, pages 435–448, New York, 1999. Marcel Dekker.

[de Amo *et al.*, 2002] S. de Amo, W. A. Carnielli, and J. Marcos. A logical framework for integrating inconsistent information in multiple databases. In T. Eiter and K.-D. Schewe, editors, *Proceedings of the* II *Symposium on Foundations of Information and Knowledge Systems* (FOIKS 2002), volume 2284 of *Lecture Notes in Computer Science*, pages 67–84, Berlin, 2002. Springer-Verlag. Preprint available at *CLE e-Prints*, 1(9), 2001.
URL = `http://www.cle.unicamp.br/e-prints/abstract_9.htm`.

[D'Ottaviano, 1990] I. M. L. D'Ottaviano. On the development of paraconsistent logic and da Costa's work. *The Journal of Non-Classical Logic*, 7(1/2):89–152, 1990.

[D'Ottaviano and da Costa, 1970] I. M. L. D'Ottaviano and N. C. A. da Costa. Sur un problème de Jaśkowski. *Comptes Rendus de l'Académie de Sciences de Paris* (A-B), 270:1349–1353, 1970.

[Došen, 1986] K. Došen. Negation as a modal operator. *Reports on Mathematical Logic*, 20:15–28, 1986.

[Epstein, 2000] R. L. Epstein. *Propositional Logics: The semantic foundations of logic,* with the assistance and collaboration of W. A. Carnielli, I. M. L. D'Ottaviano, S. Krajewski, and R. D. Maddux. Wadsworth-Thomson Learning, 2nd Edition, 2000.

[Jaśkowski, 1948] S. Jaśkowski. Rachunek zdań dla systemów dedukcyjnych sprzecznych. *Studia Societatis Scientiarun Torunesis*, Sectio A, I(5):57–77, 1948. Translated as

'A propositional calculus for inconsistent deductive systems' in *Logic and Logic Philosophy*, 7:35–56, 1999, Proceedings of the Stanisław Jaśkowski's Memorial Symposium, held in Toruń, Poland, July 1998.

[Jaśkowski, 1949] S. Jaśkowski. O koniunkcji dyskusyjnej w rachunku zdań dla systemów dedukcyjnych sprzecznych. *Studia Societatis Scientiarun Torunesis*, Sectio A, I(8):171–172, 1949. Translated as 'On the discussive conjunction in the propositional calculus for inconsistent deductive systems' in *Logic and Logic Philosophy*, 7:57–59, 1999, Proceedings of the Stanislaw Jaśkowski's Memorial Symposium, held in Toruń, Poland, July 1998.

[Johánsson, 1936] I. Johánsson. Der Minimalkalkül, ein reduzierter intuitionistischer Formalismus. *Compositio Mathematica*, 4(1):119–136, 1936.

[Kolmogorov, 1967] A. N. Kolmogorov. On the principle of excluded middle. In J. Van Heijenoort, editor, *From Frege to Gödel*, pages 414–437, Cambridge, 1967. Harvard University Press. Translation from the Russian original (1925).

[Lenzen, 1998] W. Lenzen. Necessary conditions for negation-operators (with particular applications to paraconsistent negation). In P. Besnard and A. Hunter, editors, *Reasoning with Actual and Potential Contradictions*, pages 211–239, Dordrecht, 1998. Kluwer.

[Lewin *et al.*, 1990] R. A. Lewin, I. F. Mikenberg, and M. G. Schwarze. Algebraization of paraconsistent logic $\mathbf{P}^1$. *The Journal of Non-Classical Logic*, 7(1/2):79–88, 1990.

[Loparić and da Costa, 1984] A. Loparić and N. C. A. da Costa. Paraconsistency, paracompleteness, and valuations. *Logique et Analyse (N.S.)*, 106:119–131, 1984.

[Marcos, 1999] J. Marcos. Possible-Translations Semantics (in Portuguese). Master's thesis, State University of Campinas (UNICAMP), Campinas, 1999.
URL = `http://www.cle.unicamp.br/pub/thesis/J.Marcos/`.

[Marcos, 2000] J. Marcos. 8K solutions and semi-solutions to a problem of da Costa. Draft, 2000.

[Marcos, 2004] J. Marcos. Possible-translations semantics. In W. A. Carnielli, F. M. Dionísio, and P. Mateus, editors, *Proceedings of the Workshop on Combination of Logics: Theory and applications* (CombLog'04), held in Lisbon, PT, 28–30 July 2004, pages 119–128. Departamento de Matemática, Instituto Superior Técnico, 2004. Preprint available at
URL = `http://wslc.math.ist.utl.pt/ftp/pub/MarcosJ/04-M-pts.pdf`.

[Marcos, 2005] J. Marcos. *Logics of Formal Inconsistency.* Brazil: Fundação Biblioteca Nacional, 2005 . Available at
URL = `http://wslc.math.ist.utl.pt/ftp/pub/MarcosJ/05-M-PhDthesis.pdf`.

[Marcos, 2005a] J. Marcos. Logics of essence and accident. *Bulletin of the Section of Logic*, 34(1):43–56, 2005.

[Marcos, 2005b] J. Marcos. Modality and paraconsistency. In M. Bilkova and L. Behounek, editors, *The Logica Yearbook 2004*, Proceedings of the XVIII International Symposium promoted by the Institute of Philosophy of the Academy of Sciences of the Czech Republic, held in Hejnice, CZ, 22–25 June 2004, pages 213–222. Prague: Filosofia, 2005. Preprint available at
URL = `http://wslc.math.ist.utl.pt/ftp/pub/MarcosJ/04-M-ModPar.pdf`.

[Marcos, 2005c] J. Marcos. On negation: Pure local rules. *Journal of Applied Logic*, 3(1):185–219, 2005.

[Marcos, 2005d] J. Marcos. On a problem of da Costa. In G. Sica, editor, *Essays on the Foundations of Mathematics and Logic*, volume 2, pages 39–55. Monza: Polimetrica, 2005. Reprint available at
URL = `http://www.cle.unicamp.br/e-prints/abstract_8.htm`.

[Marcos, 2005e] J. Marcos. Nearly every normal modal logic is paranormal. *Logique et Analyse (N.S.)*, 48(189/192):279–300, 2005. Preprint available at
URL = `http://wslc.math.ist.utl.pt/ftp/pub/MarcosJ/04-M-Paranormal.pdf`.

[Marcos, 2005f] J. Marcos. Possible-translations semantics for some weak classically-based paraconsistent logics. *Journal of Applied Non-Classical Logics*, in print. Preprint available at
URL = `http://www.cs.math.ist.utl.pt/ftp/pub/MarcosJ/04-M-PTS4swcbPL.pdf`.

[Marcos, 2007a] J. Marcos. Ineffable inconsistencies. In Béziau et al. [2007].

[Marcos, 2007b] J. Marcos. What is a non-truth-functional logic? Forthcoming.

[Mendelson, 1997] E. Mendelson. *Introduction to Mathematical Logic*. International Thomson Publishing, 4th Edition, 1997.

[Miller, 2000] D. Miller. Paraconsistent logic for falsificationists. In *Proceedings of the* I *Workshop on Logic and Language (Universidad de Sevilla)*, pages 197–204, Sevilla, 2000. Editorial Kronos S.A.

[Mortensen, 1980] C. Mortensen. Every quotient algebra for C_1 is trivial. *Notre Dame Journal of Formal Logic*, 21(4):694–700, 1980.

[Mortensen, 1995] C. Mortensen. *Inconsistent Mathematics,* with contributions by P. Lavers, W. James, and J. Cole. Dordrecht: Kluwer, 1995.

[Nelson, 1959] D. Nelson. Negation and separation of concepts in constructive systems. In A. Heyting, editor, *Constructivity in Mathematics,* Proceedings of the Colloquium held in Amsterdam, NL, 1957, Studies in Logic and the Foundations of Mathematics, pages 208–225. Amsterdam: North-Holland, 1959.

[Odintsov, 2005] S. Odintsov. On the structure of paraconsistent extensions of Johánsson's logic. *Journal of Applied Logic*, 3(1):43–65, 2005.

[Popper, 1948] K. R. Popper. On the theory of deduction. Parts I and II. *Indagationes Mathematicae*, 10:173–183/322–331, 1948.

[Popper, 1959] K. R. Popper. *The Logic of Scientific Discovery*. Hutchinson & Co., Ltd., London, 1959. (English translation of *Logik der Forschung*, Julius Springer Verlag, Vienna, 1936).

[Popper, 1989] K. R. Popper. *Conjectures and Refutations. The Growth of Scientific Knowledge*. Routledge & Kegan Paul, London, 5th Edition, 1989.

[Priest, 1979] G. Priest. The logic of paradox. *Journal of Philosophical Logic*, 8(2):219–241, 1979.

[Priest, 2002] G. Priest. Paraconsistent logic. In D. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic*, 2nd Edition, volume 6, pages 259–358. Kluwer Academic Publishers, 2002.

[Priest and Routley, 1989] G. Priest and R. Routley. Systems of paraconsistent logic. In Priest et al. [1989], pages 151–186.

[Priest *et al.*, 1989] G. Priest, R. Sylvan, and J. Norman, editors. *Paraconsistent Logic: Essays on the Inconsistent*. Philosophia Verlag, 1989.

[Routley and Meyer, 1976] R. Routley and R. K. Meyer. Dialectical logic, classical logic and the consistence of the world. *Studies in Soviet Thought*, 16:1–25, 1976.

[Schütte, 1960] K. Schütte. *Beweistheorie*. Springer-Verlag, Berlin, 1960.

[Seoane and de Alcantara, 1991] J. Seoane and L. P. de Alcantara. On da Costa algebras. *The Journal of Non-Classical Logic*, 8(2):41–66, 1991.

[Sette, 1973] A. M. Sette. On the propositional calculus $\mathbf{P}^1$. *Mathematica Japonicae*, 18(13):173–180, 1973.

[Shramko, 2005] Y. Shramko. Dual intuitionistic logic and a variety of negations: The logic of scientific research. *Studia Logica*, 80(2–3):347–367, 2005.

[Slater, 1995] B. H. Slater. Paraconsistent logics? *Journal of Philosophical Logic*, 24(4):451–454, 1995.

[Sylvan, 1990] R. Sylvan. Variations on da Costa **C**-systems and dual-intuitionistic logics. I. Analyses of C_ω and CC_ω. *Studia Logica*, 49(1):47–65, 1990.

[Urbas, 1989] I. Urbas. Paraconsistency and the **C**-systems of da Costa. *Notre Dame Journal of Formal Logic*, 30(4):583–597, 1989.

[Urbas, 1990] I. Urbas. Paraconsistency. *Studies in Soviet Thought*, 39:343–354, 1990.

[Urbas, 1996] I. Urbas. Dual-intuitionistic logic. *Notre Dame Journal of Formal Logic*, 37(3):440–451, 1996.

[Vakarelov, 1989] D. Vakarelov. Consistency, completeness and negation. In Priest et al. [1989], pages 328–363.

[Wójcicki, 1988] R. Wójcicki. *Theory of Logical Calculi*. Synthese Library. Kluwer Academic Publishers, 1988.

Walter Carnielli
Department of Philosophy, IFCH, and
Centre for Logic, Epistemology and The History of Science (CLE)
State University of Campinas (UNICAMP), Brazil
Security and Quantum Information Group (SQIG), IST / IT, Portugal

Marcelo E. Coniglio
Department of Philosophy, IFCH, and
Centre for Logic, Epistemology and The History of Science (CLE)
State University of Campinas (UNICAMP), Brazil
Security and Quantum Information Group (SQIG), IST / IT, Portugal

João Marcos
Department of Informatics and Applied Mathematics (DIMAp), CCET,
Federal University of Rio Grande do Norte (UFRN), Brazil
Security and Quantum Information Group (SQIG), IST / IT, Portugal

JON WILLIAMSON

CAUSALITY

1 INTRODUCTION

Perhaps the key philosophical questions concerning causality are the following:

- what are causal relationships?
- how can one discover causal relationships?
- how should one reason with causal relationships?

This chapter will focus on the first two questions. The last question is equally important — of course we need to know the best way to make predictions, perform diagnoses and make strategic decisions — but in the absence of a well-entrenched mathematical calculus of causality, the answers given to the last question tend to depend on the answers provided to the first two questions.

Standard responses to the first, ontological question are surveyed in §2, while the second, epistemological question is dealt with in §3. I advocate a position I call *epistemic causality* which is sketched in §4, and which is compared to the positions of Judea Pearl in §5 and Huw Price in §6.[1]

2 THE NATURE OF CAUSALITY

There are three varieties of position on causality. One can argue that the concept of causality is of heuristic use only and should be eliminated from scientific discourse: this was the tack pursued by Bertrand Russell, who maintained that science appeals to functional relationships rather than causal laws.[2] Alternatively one can argue that causality is a fundamental feature of the world and should be treated as a scientific primitive — this claim is usually the result of disillusionment with purported philosophical analyses, several of which appeal to the asymmetry of time in order to explain the asymmetry of causation, a strategy that is unattractive to those who want to analyse time in terms of causality. Or one can maintain

[1] Epistemic causality motivates an answer to the last question, how should one reason with causal relationships? The ensuing formalism is presented in detail in [Williamson, 2004].

[2] [Russell, 1913]. Russell later modified his views on causality, becoming more tolerant of the notion.

D.M. Gabbay and F. Guenthner (eds.), Handbook of Philosophical Logic, 2nd Edition, Volume 14, 95–126.

that causal relations can be reduced to other concepts not involving causal notions. This latter position is dominant in the philosophical literature, and there are four main approaches which can be described roughly as follows. The mechanistic theory, discussed in §2.1, reduces causal relations to physical processes. The probabilistic account (§2.2) reduces causal relations to probabilistic relations. The counterfactual account (§2.3) reduces causal relations to counterfactual conditionals. The agent-oriented account (§2.4) reduces causal relations to the ability of agents to achieve goals by manipulating their causes.[3]

2.1 Mechanisms

The mechanistic account of causality aims to understand the physical processes that link cause and effect, interpreting causal statements as saying something about such processes. Proponents of this type of position include Wesley Salmon[4] and Phil Dowe.[5] They argue that a causal process is one that transmits[6] or possesses[7] a conserved physical quantity, such as energy-mass, linear momentum or charge, from start (cause) to finish (effect).

The mechanistic account is clearly a physical interpretation of causality, since it identifies causal relationships with physical processes. Such a notion of cause relates single cases, since only they are linked by physical processes, although causal regularities or laws may be induced from single-case causal connections.

The main limitation of this approach is its rather narrow applicability: most of our causal assertions are apparently unrelated to the physics of conserved quantities. While it may be possible that physical processes such as those along which quantities are conserved could suggest causal links to physicists, such processes are altogether too low-level to suggest causal relationships in economics, for instance. One could maintain that the economists' concept of causality is the same as that of physics and is reducible to physical processes,[8] but one would be forced to accept that the epistemology of such a concept is totally unrelated to its metaphysics. This is undesirable: if the grounds for knowledge of a causal connection have little to do with the nature of the causal connection as it is analysed then one can argue that it cannot be the causal connection that we have knowledge

[3] See the introduction to [Sosa and Tooley, 1993] for more discussion on the variety of interpretations of causality.

[4] [Salmon, 1980], [Salmon, 1984], [Salmon, 1997], [Salmon, 1998].

[5] [Dowe, 1993], [Dowe, 1996], [Dowe, 1999], [Dowe, 2000], [Dowe, 2000b].

[6] [Salmon, 1997] §2.

[7] [Dowe, 2000b] §V.1.

[8] This was the tack Salmon took in connection with his earlier theory that conceived of causal processes as involving the transmission of marks rather than conserved quantities. See [Salmon, 1998], page 206.

of, but something else.[9] On the other hand one could keep the physical account and accept that the economists' causality differs from the physicists' causality. But this position faces the further questions of what economists' causality is, and why we think that cause is a single concept when in fact it isn't. These problems clearly motivate a more unified account of causality.

2.2 Probabilistic Causality

Probabilistic causality has a wider scope than the mechanistic approach: here the idea is to understand causal connections in terms of probabilistic relationships between variables, be they variables in physics, economics or wherever. There is no firm consensus among proponents of probabilistic causality as to what probabilistic relationships among variables constitute causal relationships, but typically they appeal to the intuitions behind the *Principle of the Common Cause*: if two variables are probabilistically dependent then one causes the other or they are effects of common causes which screen off the dependence (i.e. the two variables are probabilistically independent conditional on the common causes). Indeed Hans Reichenbach applied the Principle of the Common Cause to an analysis of causality, as a step on the way to a probabilistic analysis of the direction of time.[10] Similarly Patrick Suppes argued that causal relations induce probabilistic dependencies and that screening off can be used to differentiate between variables that are common effects and variables that are cause and effect.[11] However, both these analyses fell foul of a number of criticisms,[12] and more recent probabilistic approaches adopt *Causal Dependence* (cause and direct effect are probabilistically dependent conditional on the effect's other direct causes) and the *Causal Markov Condition* (each variable is probabilistically independent of its non-effects, conditional on its direct causes) as necessary conditions for causality, together with other less central conditions which are sketched in §3.[13] Sometimes Causal Dependence is only implicitly adopted: the causal relation may be defined as the smallest relation that satisfies the Causal Markov Condition, in which case Causal Dependence must hold.

Probabilistic causality is normally applied to repeatably-instantiatable rather than single-case variables — in principle either is possible, as long as the chosen interpretation of probability handles the same kind of variables. Invariably causality is interpreted as a physical, mind-independent concept.

[9] See [Benacerraf, 1973] for a parallel argument in mathematics.

[10] [Reichenbach, 1956].

[11] [Suppes, 1970].

[12] See [Salmon, 1980b], §§2-3

[13] See [Pearl, 1988], [Pearl, 2000], [Spirtes *et al.*, 1993], [McKim and Turner, 1997] and [Korb, 1999]. Note that the concept of *direct* cause does not require that causal chains be discrete. It is merely presumed that Causal Dependence or the Causal Markov Condition will hold where the direct causes are taken to be a set of causes that are sufficiently close to the effect, with one direct cause per causal chain that leads to the effect.

The chief problem that besets probabilistic causality is the dubious status of the probabilistic conditions to which the account appeals. While the conditions seem intuitive and might be expected to hold much of the time there are clear cases where they fail. The Principle of the Common Cause and the Causal Markov Condition are widely acknowledged to fail in certain cases that crop up in quantum mechanics, but they also fail more generally wherever probabilistic dependencies are induced by non-causal relationships: where variables are semantically, logically or mathematically related, or they are related by non-causal physical laws (as in the quantum mechanics case) or boundary conditions.[14] Causal Dependence fails for instance where an event must be caused by one of two equally efficacious physical processes: if a machine can be activated by precisely one of two fully reliable power supplies, then the choice of power supply will not change the probability of its direct effect, the machine being activated.[15] Of course it is not good enough for a probabilistic analysis of causality if the defining connection between probability and causality admits exceptions — we are left with the question as to how causality is to be analysed in the exceptional cases.

2.3 Counterfactuals

The counterfactual account, developed in detail by David Lewis,[16] reduces causal relations to subjunctive conditionals: C is a direct cause of E if and only if (i) if C were to occur then E would occur (or its chance of occurring would be significantly raised) and (ii) if C were not to occur then E would not occur (or its chance of occurring would be significantly lowered). The subjunctive conditionals (called *counterfactual* conditionals if the antecedent is false) are in turn given a semantics in terms of possible worlds: 'if C were to occur then E would occur' is true if and only if (i) there are no possible worlds in which C is true or (ii) E holds at all the possible worlds in which C holds that our closest to our own world. So causal claims are claims about what goes on in possible worlds that are close to our own.[17]

Lewis's counterfactual theory was developed to account for causal relationships between single-case events (which can be thought of as single-case variables which take the values 'occurs' or 'does not occur'), and the causal relation is intended to be mind-independent and objective.

Many of the difficulties with this view stem from Lewis' reliance on possible worlds. Possible worlds are not just a dispensable façon de parler for

[14] These counterexamples are explained in detail in [Williamson, 2004], §4.2.

[15] [Williamson, 2004] §7.3.

[16] [Lewis, 1973].

[17] Lewis modified his account in [Lewis, 2000], but the changes made have little bearing on our discussion. See [Lewis, 1986b] for Lewis' account of causal explanation.

Lewis, they are assumed to exist in just the way our world exists. But we have no physical contact with these other worlds, which makes it hard to see how their goings-on can be the object of our causal claims and hard to see how we discover causal relationships. Moreover it is doubtful whether there is an objective way to determine which worlds are closest to our own if we follow Lewis' suggestion of measuring closeness by similarity — two worlds are similar in some respects and different in others and choice or weighting of these respects is a subjective matter. Causal relations, on the other hand, do not seem to be subjective. Instead of analysing causal relations, of which we have at least an intuitive grasp, in terms of subjunctive conditionals and ultimately possible worlds, which many find mysterious, it would be more natural to proceed in the opposite direction. Thus we might be better-off appealing to causality to decide whether E would (be more likely to) occur were C to occur,[18] and depending on the answer we could then say whether a world in which C and E occurs is closer to our own than one in which C occurs but E does not.

2.4 Agency

The agency account, whose chief recent proponents are perhaps Huw Price and Peter Menzies,[19] analyses causal relations in terms of the ability of agents to achieve goals by manipulating their causes. According to this account, C causes E if and only if bringing about C would be an effective way for an agent to bring about E. Here the strategy of bringing about C is deemed effective if a rational decision theory would prescribe it as a way of bringing about E. Menzies and Price argue that the strategy would be prescribed if and only if it raises the 'agent probability' of the occurrence of E.[20] (The events they consider are single-case.)

Menzies and Price do not agree as to the interpretation of these probabilities: Menzies maintains that they are chances, while Price seems to have a Bayesian conception.[21] Consequently it is not entirely clear whether they view causality as a physical or mental notion. On the one hand they claim that there would be causal relations without agents,[22] while on the other they say, 'we would argue that when an agent can bring about one event as a means to bringing about another, this is true in virtue of certain basic intrinsic features of the situation involved, these features being essentially

[18] See [Pearl, 2000], chapter 7, for an analysis of counterfactuals in terms of causal relations. [Dawid, 2001] argues that counterfactuals are irrelevant and misleading for an analysis of causality.

[19] [Price, 1991], [Price, 1992], [Price, 1992b], [Menzies and Price, 1993].

[20] [Menzies and Price, 1993].

[21] [Menzies and Price, 1993] pg. 190.

[22] [Menzies and Price, 1993] §6.

non-causal though not necessarily physical in character',[23] and maintain that the concept of cause is a 'secondary quality', relative to human responses or capacities.[24] From this relativity one might expect cause to be subjective, but they say that causation is significantly more objective than other secondary quantities like colour or taste.[25] We shall examine Price's views on these matters in more detail in §6.

The main problems that beset the agency approach are inherited from those faced by the probabilistic and counterfactual approaches. First, the agency approach assumes a version of Causal Dependence for agent probabilities — we saw in §2.2 that this condition does not always hold.[26] Of course, where a causal connection is not accompanied by probabilistic dependence, such as in the power supply example of §2.2, bringing about a cause is not a good strategy for bringing about its effects. Second, the agency account appeals to subjunctive conditionals[27] (C causes E if and only if, *were* an agent to bring about C, that *would* be a good strategy for bringing about E) and so qualms about the utility of a counterfactual account can equally be applied to the agency approach.

3 DISCOVERING CAUSAL RELATIONSHIPS

Different views on the nature of causality lead to different suggestions for discovering causal relationships. The mechanistic view of causality, for example, leads naturally to a quest for physical processes, while proponents of probabilistic causality prescribe searching for probabilistic dependencies and independencies.

However there are two very general strategies for causal discovery which cut across the ontological positions. Whatever view one holds on the nature of causality, one can advocate either *hypothetico-deductive* or *inductive* discovery of causal relationships. Under a hypothetico-deductive account (§3.1) one hypothesises causal relationships, deduces predictions from the hypothesis, and then tests the hypothesis by seeing how well the predictions accord with what actually happens. Under an inductive account (§3.2), one makes a large number of observations and induces causal relationships directly from this mass of data. We shall discuss each of these approaches

[23] [Menzies and Price, 1993] pg. 197.
[24] [Menzies and Price, 1993] pp. 188,199.
[25] [Menzies and Price, 1993] pg. 200.
[26] In fact the version assumed by the agency approach does not restrict attention to direct causes and does not demand that dependence be conditional on the effect's other causes. This type of dependence condition is rarely advocated since it faces a wider range of counterexamples than Causal Dependence in the form used here — see the references given in §2.2.
[27] [Menzies and Price, 1993] §5.

in turn in this chapter, and give an overview of some recent proposals for discovering causal relationships.

3.1 Hypothetico-Deductive Discovery

According to the hypothetico-deductive account, a scientist first hypothesises causal relationships and then tests this hypothesis by seeing whether predictions drawn from it are borne out. The testing phase may be influenced by views on the nature of causality: a causal hypothesis can be supported or refuted according to whether physical processes are found that underlie the hypothesised causal relationships, whether probabilistic consequences of the hypothesis are verified, and whether experiments show that by manipulating the hypothesised causes one can achieve their effects.

Karl Popper was an exponent of the hypothetico-deductive approach. For Popper a causal explanation of an event consists of natural laws (which are universal statements) together with initial conditions (which are single-case statements) from which one can predict by deduction the event to be explained. The initial conditions are called the 'cause' of the event to be explained, which is in turn called the 'effect'.[28] Causal laws, then, are just universal laws, and are to be discovered via Popper's general scheme for scientific discovery: (i) hypothesise the laws; (ii) deduce their consequences, rejecting the laws and returning to step (i) if these consequences are falsified by evidence. Popper thus combines what is known as the *covering-law* account of causal explanation with a hypothetico-deductive account of learning causal relationships.

The covering-law model of explanation was developed by Hempel and Oppenheim[29] and also Railton,[30] and criticised by Lewis.[31] While such a model fits well with Popper's general account of scientific discovery, neither the details nor the viability of the covering-law model are relevant to the issue at stake: a Popperian hypothetico-deductive account of causal discovery can be combined with practically any account of causality and causal explanation.[32] Neither does one have to be a strict falsificationist to adopt a hypothetico-deductive account. Popper argued that the testing of a law only proceeds by falsification: a law should be rejected if contradicted by observed evidence (i.e. if falsified), but should never be accepted or regarded as confirmed in the absence of a falsification. This second claim of Popper's

[28] [Popper, 1934] §12.
[29] [Hempel and Oppenheim, 1948].
[30] [Railton, 1978].
[31] [Lewis, 1986b] §VII.
[32] Even Russell's eliminativist position of [Russell, 1913], in which he argued that talk of causal laws should be eradicated in favour of talk of functional relationships, ties in well with Popper's logic of scientific discovery. Both Popper and Russell, after all, drew no sharp distinction between causal laws and the other universal laws that feature in science.

has often been disputed, and many argue that a hypothesis is confirmed by evidence in proportion to the probability of the hypothesis conditional on the evidence.[33] Given this probabilistic measure of confirmation — or indeed any other measure — one can accept the hypothesised causal relationships according to the extent to which evidence confirms the hypothesis. Thus the hypothetico-deductive strategy for learning causal relationships is very general: it does not require any particular metaphysics of causality, nor a covering-law model of causal explanation, nor a strict falsificationist account of testing.

Besides providing some criterion for accepting or rejecting hypothesised causal relationships, the proponent of a hypothetico-deductive account must do two things: (i) say how causal relationships are to be hypothesised; (ii) say how predictions are to be deduced from the causal relationships.

Popper fulfilled the latter task straightforwardly: effects are predicted as logical consequences of laws given causes (initial conditions). The viability of this response hinges very closely on Popper's account of causal explanation, and the response is ultimately inadequate for the simple reason that no one accepts the covering-law model as Popper formulated it: more recent covering-law models are significantly more complex, coping with chance explanations.[34]

Popper's response to the former task was equally straightforward, but perhaps even less satisfying:

> my view of the matter, for what it is worth, is that there is no such thing as a logical method of having new ideas, or a logical reconstruction of this process. My view may be expressed by saying that every discovery contains 'an irrational element', or 'a creative intuition'[35]

Popper accordingly placed the question of discovery firmly in the hands of psychologists, and concentrated solely on the question of the justification of a hypothesis.

The difficulty here is that while hypothesising may contain an irrational element, Popper has failed to shed any light on the rational element which must surely play a significant role in discovery. Popper's scepticism about the existence of a logic need not have precluded any discussion of the act of hypothesising from a normative point of view: both Popper in science and Pólya in mathematics remained pessimistic about the existence of a precise logic for hypothesising, yet Pólya managed to identify several imprecise but important heuristics.[36] One particular problem is this: a theory may be refuted by one experiment but perform well in many others; in such

[33] See [Howson and Urbach, 1989], [Earman, 1992].
[34] [Railton, 1978] for example.
[35] [Popper, 1934] pg. 32.
[36] [Polya, 1945], [Polya, 1954], [Polya, 1954b].

a case it may need only some local revision, to deal with the domain of application on which it is refuted, rather than wholesale rehypothesising. Popper's account says nothing of this, giving the impression that with each refutation one must return to a blank sheet and hypothesise afresh. The hypothetico-deductive method as stated neither gives an account of the progress of scientific theories in general, nor of causal theories in particular.

Any hypothetico-deductive account of causal discovery which fails to probe either the hypothetico or the deductive aspects of the process is clearly lacking. These are, in my view, the key shortcomings of Popper's position. I shall try to shed some light on these aspects when I present a new type of hypothetico-deductive account in §4.5. For now, we shall turn to a competing account of causal discovery, inductivism.

3.2 Inductive Learning

Francis Bacon developed a rather different account of scientific learning. First one makes a large amount of careful observations of the phenomenon to be explained, by performing experiments if need be. One compiles a table of positive instances (cases in which the phenomenon occurs),[37] a table of negative instances (cases in which the phenomenon does not occur)[38] and a table of partial instances (cases in which the phenomenon occurs to a certain degree).[39]

> We have chosen to call the task and function of these three tables the *Presentation of instances to the intellect.* After the *presentation* has been made, *induction* itself has to be put to work. For in addition to the *presentation* of each and every instance, we have to discover which nature appears constantly with a given nature or not, which grows with it or decreases with it; and which is a limitation (as we said above) of a more general nature. If the mind attempts to do this affirmatively from the beginning (as it always does if left to itself), fancies will arise and conjectures and poorly defined notions and axioms needing daily correction, unless one chooses (in the manner of the Schoolmen) to defend the indefensible.[40]

Thus Bacon's method consists of presentation followed by induction of a theory from the observations. It is to be preferred over a hypothetico-deductive approach because it avoids the construction of poor hypotheses in the absence of observations, and it avoids the tendency to defend the indefensible:

[37] [Bacon, 1620] §II.XI.
[38] [Bacon, 1620] §II.XII.
[39] [Bacon, 1620] §II.XIII.
[40] [Bacon, 1620] §II.XV.

> Once a man's understanding has settled on something (either because it is an accepted belief or because it pleases him), it draws everything else also to support and agree with it. And if it encounters a larger number of more powerful countervailing examples, it either fails to notice them, or disregards them, or makes fine distinctions to dismiss and reject them, and all this with much dangerous prejudice, to preserve the authority of its first conceptions.[41]

Note that while Bacon's position is antithetical to Popper's hypothetico-deductive approach, it is compatible with Popper's falsificationism — indeed Bacon claims that 'every *contradictory instance* destroys a conjecture'.[42] The first step of the inductive process, *exclusion*, involves ruling out a selection of simple and often rather vaguely formulated conjectures by means of providing contradictory instances.[43] The next step is a *first harvest*, which is a preliminary interpretation of the phenomenon of interest.[44] Bacon then produces a seven-stage process of elucidating, refining and testing this interpretation — only the first stage of which was worked out in any detail.[45]

Present-day inductivists claim that causal relationships can be inferred algorithmically from experimental and observational data, and that suitable data would yield the correct causal relationships. Usually, but not necessarily, the data takes the form of a database of past cases: a set V of repeatably instantiatable variables are measured, each entry of the database $D = (u_1, \ldots, u_k)$ consists of an observed assignment of values to some subset U_i of V. Such an account of learning is occasionally alluded to in connection with probabilistic analyses of causality and has been systematically investigated by researchers in the field of artificial intelligence, including groups in Pittsburgh,[46] Los Angeles[47] and Monash,[48] proponents of a Bayesian learning approach,[49] and computationally-minded psychologists.[50]

These approaches seek to learn various types of causal model. The simplest type of causal model is just a *causal graph* (i.e. a directed acyclic graph in which nodes correspond to variables and there is an arrow from one node

[41] [Bacon, 1620] §I.XLVI.
[42] [Bacon, 1620] §II.XVIII.
[43] [Bacon, 1620] §§II.XVIII-XIX.
[44] [Bacon, 1620] §II.XX.
[45] [Bacon, 1620] §§II.XXI-LII.
[46] [Spirtes *et al.*, 1993], [Scheines, 1997], [Glymour, 1997], [Mani and Cooper, 1999], [Mani and Cooper, 2000], [Mani and Cooper, 2001].
[47] [Pearl, 2000], [Pearl, 1999].
[48] [Dai *et al.*, 1997], [Wallace and Korb, 1999], [Korb and Nicholson, 2003].
[49] [Heckerman *et al.*, 1999], [Cooper, 1999], [Cooper, 2000], [Tong and Koller, 2001], [Yoo *et al.*, 2002].
[50] [Waldmann and Martignon, 1998], [Waldmann, 2001], [Tenenbaum and Griffiths, 2001], [Glymour, 2001], [Hagmayer and Waldmann, 2002].

to another if the former directly causes the latter) which shows only qualitative causal relationships. A *causal net* is slightly more complex, containing not only a qualitative causal graph but also quantitative information, the probability distribution $p(a_i|par_i)$ of each variable A_i conditional on its parents Par_i, the direct causes of A_i in the graph. A *structural equation model* is a third type of causal model — this can be thought of as a causal graph together with an equation for each variable in terms of its direct cause variables, $A_i = f_i(Par_i, E_i)$, where f_i is some function and E_i is an error variable.

The mainstream of these inductivist AI approaches have the following feature in common. In order that causal relationships can be gleaned from statistical relationships, the approaches assume the Causal Markov Condition.[51] A causal net contains the Causal Markov Condition as an inbuilt assumption; in the case of structural equation models the Causal Markov Condition is a consequence of the representation of each variable as a function just of its direct causes and an error variable, given the further assumption that all error variables are probabilistically independent.

The inductive procedure then consists in finding the class of causal models — or under some approaches a single 'best' causal model — whose probabilistic independencies implied via the Causal Markov Condition are consistent with independencies inferred from the data. Other assumptions are often also made, such as minimality (no submodel of the causal model also satisfies the Causal Markov Condition), faithfulness (all independencies in the data are implied via the Causal Markov Condition), linearity (all variables are linear functions of their direct causes and uncorrelated error variables), causal sufficiency (all common causes of measured variables are measured), context generality (every individual possesses the causal relations of the population), no side effects (one can intervene to fix the value of a variable without changing the value of any non-effects of the variable) and determinism. However these extra assumptions are less central than the Causal Markov Condition: approaches differ as to which of these extra assumptions they adopt and the assumptions tend to be used just to facilitate the inductive procedure based on the Causal Markov Condition, either by helping to provide some justification of the inductive procedure or by increasing the purported efficiency or efficacy of algorithms for causal induction.[52]

The brunt of criticism of the inductive approach tends to focus on the Causal Markov Condition and the ancillary assumptions outlined above. I

[51] There are inductive AI methods that take a totally different approach to causal learning, such as that in [Karimi and Hamilton, 2000] and [Karimi and Hamilton, 2001], and [Wendelken and Shastri, 2000]. However, non-Causal-Markov approaches are well in the minority.

[52] See Chapter 8 of [Williamson, 2004] for a more detailed overview of inductive algorithms for causal discovery.

have already mentioned the difficulties that beset the Causal Markov Condition; in cases where this condition fails the inductive approach will simply posit the wrong causal relationships. It is plain to see that the ancillary conditions are also very strong and these face numerous counterexamples themselves. The proof, inductivists claim, will be in the pudding. However, the reported successes of inductive methods have been questioned,[53] and these criticisms lend further doubt to the inductive approach as a whole and the Causal Markov Condition in particular as its central assumption.[54]

Unfortunately neither Popper's hypothetico-deductive approach nor the recent inductivist proposals from AI offer a viable account of the discovery of causal relationships. Popper's hypothetico-deductive approach suffers from underspecification: the hypothesis of causal relationships remains a mystery and Popper's proposals for deducing predictions from hypotheses were woefully simplistic. On the other hand, the key shortcoming of the inductive approach is this: given the counterexamples to the Causal Markov Condition the inductive approach cannot guarantee that the induced causal model or class of causal models will tally with causality as we understand it — the causal models that result from the inductive approach will satisfy the Causal Markov Condition, but the true causal picture may not. While this objection may put paid to the dream of using Causal Markov formalisms for learning causal relationships, an alternative formalism may yet ground the inductive approach. In §4.5 we shall see that the inductive and hypothetico-deductive approaches can be reconciled by using new inductive methods as a way of hypothesising a causal model, then deducing its consequences and restructuring the model if these are not borne out.

4 EPISTEMIC CAUSALITY

In this section I shall sketch my own view of causality, *epistemic causality*. A more detailed exposition can be found in [Williamson, 2004].

As I see it, current theories of causality suffer from over-compartmentalisation. Current theories analyse causality in terms of just one of the indicators of causal relationships — mechanisms, probabilistic dependencies or independencies, counterfactuals or agency considerations — to the expense of the others. While one indicator may be more closely connected with causality than the others, our causal beliefs are clearly based on several indicators, not exclusively on one. It seems that if we are to understand

[53] [Humphreys and Freedman, 1996], [Humphreys, 1997], [Freedman and Humphreys, 1999], [Woodward, 1997].

[54] See [Dash and Druzdzel, 1999], [Hausman, 1999], [Hausman and Woodward, 1999], Part Three of [Glymour and Cooper, 1999], [Lemmer, 1996], [Lad, 1999], [Cartwright, 1997], [Cartwright, 1999] and [Cartwright, 2001] for further discussion of the inductive approach.

the complexity of causality we must focus on our causal beliefs and the role these indicators have in forming them.

Epistemic causality focusses on causal beliefs. It provides an account of causal beliefs in informal causal reasoning (§4.1), as well as a more formal account of how we ought to determine causal beliefs (§4.2). It takes causality to be an objective notion (§4.3) yet primarily a mental construct (§4.4). And it provides an account of the discovery of causal relationships (§4.5).

4.1 Informal Causal Reasoning

Why do we have causal beliefs? The answer to this fundamental question, according to the epistemic view, is based on the following doctrines:

Convenience It is convenient to represent the world in terms of cause and effect.

Explanation Humans think in terms of cause and effect because of this convenience, not because there is something physical corresponding to cause which humans experience.

It is convenient to represent the world in terms of cause and effect because a causal representation, if correct, enables us to make successful causal inferences: it allows us to make correct predictions, correct diagnoses and successful strategic decisions. Correct predictions and diagnoses are possible since, typically, cause and direct effect are probabilistically dependent. Successful strategic decisions are possible since, typically, manipulating a cause is a good way of changing its direct effects. (Note that here it is enough that these associations are *typical*; on the other hand an analysis of causality in terms of these associations would be flummoxed by the existence of counterexamples.)

It is clear why the convenience of causality explains our having causal beliefs: successful causal reasoning has survival value. It doesn't take us long as babies to learn that crying brings us food. The value of correctly predicting the effect of a fault in a power plant, correctly diagnosing an ulcer, or successfully manipulating the economy is equally apparent.

The Explanation thesis divorces causal beliefs from any physical, mind-independent notion of causality. While one might remain agnostic as to whether there are physical causal relationships, one might instead adopt an *anti-physical* position, claiming that in the interests of ontological parsimony one should reject physical causality. I leave the selection of an appropriate stance here entirely open.

4.2 Formal Causal Reasoning

The starting-point of a more formal account of causal beliefs is to ask how one might determine a directed acyclic causal graph $\mathcal{C}_\beta$ that depicts the

causal beliefs that an agent ought to adopt on the basis of her background knowledge β.

Arguably $\mathcal{C}_\beta$ should be compatible with background knowledge β, but should otherwise be as non-committal as possible. The agent's causal beliefs should include those causal claims warranted by her background knowledge but no unwarranted causal claims. Since each arrow in a causal graph makes a causal claim, $\mathcal{C}_\beta$ should be a graph that contains fewest arrows, from all those graphs that are compatible with β.

Thus we need to determine which graphs are compatible with background knowledge β. Given the above discussion of informal causal reasoning it seems natural to suppose that a causal graph that is compatible with β should be a good causal representation of β, in the sense that its causal claims should represent any predictive, diagnostic and strategic relationships that can be gleaned from β. We can explicate this thought by insisting that the causal graph include an arrow from A to B if:

- A and B represent non-overlapping physical events (so A and B are the kinds of things that might be causally related, rather than semantically, logically or mathematically related),[55]
- B is *strategically dependent* on A: intervening to change A can change the probability of B, when B's other direct causes are controlled for,
- this dependence is not otherwise accounted for by the agent's background knowledge or other beliefs, and
- the inclusion of this arrow is not inconsistent with other background knowledge. It is here that the other various indicators of causality get taken into account: for instance if it is known that there is no physical mechanism linking A with B, or if it is known that A only occurs after B, then the agent should not deem A to be a direct cause of B.

In sum then, the agent's causal belief graph $\mathcal{C}_\beta$ should be a graph, from all those that are compatible with β in the sense outlined above, that has fewest arrows.

Given this concept of a causal causal belief graph, it is not hard to see that the Causal Markov Condition and the Principle of the Common Cause will hold when $\mathcal{C}_\beta$ contains an arrow for each strategic dependency, and that Causal Dependence will hold if furthermore each arrow in $\mathcal{C}_\beta$ corresponds to a strategic dependency. In this latter case $\mathcal{C}_\beta$ will be a minimal graph satisfying the Causal Markov Condition. We thus have a *qualified* justification of the three controversial principles that connect causality and probability, and a *qualified* justification of inductive methods for causal learning that infer a minimal graph satisfying the Causal Markov Condition.

[55] In fact this is too strict. A causal graph can also feature as a cause or effect — see [Williamson, 2004], Chapter 10.

4.3 The Objectivity of Causality

Clearly a primary desideratum of any theory of causality is that it account for the apparent objectivity of causal notions: causal claims do not appear to be arbitrary, a matter of personal opinion. It might be thought that epistemic causality, focussing as it does on causal beliefs, suffers in this respect. On the contrary, epistemic causality leads to an objective concept of cause as we shall see now.

The word 'objectivity' is routinely used to mean many different things, but the meaning most relevant to discussions of causality is *lack of arbitrariness*. It is important that causal claims are not arbitrary in a pathological way. Note that objectivity in this sense is a matter of degree: if any set of causal claims is correct then causality is *fully subjective*; at the other end of the scale if only one set of causal claims is correct then causality is *fully objective*; degree of objectivity increases as arbitrariness, i.e. the proportion of causal claims that are correct, decreases. We shall be interested in two points on this scale:

Epistemic Objectivity If two agents with the same background knowledge disagree as to causal relationships then at least one of them must be wrong.

Full Objectivity If two agents disagree as to causal relationships then at least one of them must be wrong.

The causal belief graph $\mathcal{C}_\beta$ that an agent ought to adopt on the basis of background knowledge β is epistemically objective (rather, close to epistemically objective: there may be more than one minimal graph compatible with β, but there tends to be little room for subjectivity).

Note that epistemic objectivity is enough for the requirements of science. Sciences demand that disagreements should be resolvable on the basis of current background knowledge in the scientific literature: if there is a disagreement as to whether or not the claim that smoking causes cancer is warranted by current evidence, at least one party should be wrong, for otherwise arbitrariness would render such debates pointless.

Philosophical preconceptions require more though — something close to full objectivity. Intuitively there is a fact of the matter as to what causes what, and if indeed causality is fully objective, a theory of causality should be able to capture this characteristic. The standard way of explaining full objectivity of a scientific concept is to suppose that the concept refers to something physical and mind-independent. Then if there is disagreement as to claims about the concept, the correctness of these claims are decided on the basis of their truth when taken as claims about physical reality.

But projecting a concept onto the physical world is not the only way to account for its full objectivity. Full objectivity can also be generated

from epistemic objectivity. A (close to) fully objective causal graph $\mathcal{C}^*$ can be interpreted as $\mathcal{C}_{\beta^*}$, the causal belief graph one ought to adopt on the basis of some *ultimate background knowledge* β^*. This is the *ultimate belief* interpretation of causality.

What constitutes ultimate background knowledge? There are two possible approaches here.

One might choose β^* to be *limiting* background knowledge, to which an agent's background knowledge tends as time progresses. Now different agents' knowledge might be expected to tend to different limits, so one needs to distinguish a special agent. When C.S. Peirce wanted to analyse truth as the limit of belief, he chose science as the agent whose beliefs are privileged.[56] In our context we might take β^* to be the limit of scientific inquiry. The difficulties with this suggestion are (i) that science is not unanimous: different scientific parties and different scientific theories contradict each other, making it difficult to extract a consistent body of knowledge from science at any particular time, and (ii) that scientific knowledge is no longer considered to be accumulative: science undergoes revolutions, radical changes in scientific knowledge, and thus it is by no means clear that scientific knowledge will tend to a fixed limit. A further problem with this general strategy stems from the way it ties causality very closely to a *particular* agent (science or whomsoever): if the agent had been different, her background knowledge may have been very different, in which case her limiting beliefs and thus causality itself would be very different. This seems counter-intuitive. Under the epistemic account, a causal model is a convenient way of representing the world. While causal relations might be expected to depend on the contingencies of the world, they should not be expected to depend on non-epistemic contingencies of a particular agent.

A natural alternative strategy is to consider the characterising feature of causality, its convenience, and choose β^* that optimises the convenience of $\mathcal{C}^* = \mathcal{C}_{\beta^*}$. (This approach corresponds to William James' analysis of truth: '*The true is the name of whatever proves itself to be good in the way of belief.*'[57]) Now causal beliefs will provide the most convenient representation of the world if they are based on the fullest knowledge of the world, i.e. if β^* contains knowledge of all the indicators of causality. Thus we can take β^* to consist of knowledge of all probabilities, physical mechanisms, temporal relations, non-causal inducers of probabilistic dependencies (semantic, logical and mathematical relationships, non-causal physical laws and boundary conditions) and so on. This strategy has the advantages that β^* is well defined (as long as the indicators of causality can be delimited) and that causality is not tied to a particular agent — indeed causality is not tied even to there *being any agents*.

[56] [Peirce, 1877].
[57] [James, 1907] 30.

We see then how epistemic causality can provide the (close to epistemic) objectivity required for science and the (close to full) objectivity required to satisfy our intuitions about causality.

4.4 What Causality Is

To summarise, epistemic causality provides both an account of causal beliefs and of a fully objective notion of causality. It deals with the causal beliefs an agent ought to adopt on the basis of her background knowledge, and considers causality itself to be the causal beliefs that an agent ought to adopt on the basis of full knowledge of the indicators of causality.

In that sense causality is a mental notion, not a physical notion. This mental metaphysics for causality stands shoulder to shoulder with causal epistemology: the causal relation is just an ultimate set of causal beliefs. Moreover the anti-physical version of epistemic causality makes the further claim that this is the only notion of cause — there is no such thing as physical causality.

But causality is not mental in any degenerate psychologistic sense. Causality does not depend on the mind of any particular agent — it is a normative notion and causal relations are as mind-independent as the laws of logic. Causality is not subject to the whim of an agent: a rational agent can exercise little or no choice when she forms her causal beliefs; there is little or no arbitrariness as what the correct causal relationships are. Causality is objective.

Note that although epistemic causality can be construed as a subjunctive theory, claiming that *were* an agent to know β and *were* she rational then she *would* believe $\mathcal{C}_\beta$, it does not suffer from the problems that beset a counterfactual analysis of causality. This is because its subjunctive conditional claims are not given a semantics in terms of possible worlds — instead a theory of rational causal belief is developed to explicate their meaning. Thus worries about possible worlds do not translate into worries about the claims of epistemic causality.

4.5 Discovery of Causal Relationships

Epistemic causality breaks the barriers between the hypothetico-deductive and inductive accounts of discovering causal relationships.

On the one hand epistemic causality advocates an inductive approach to causal discovery. Given observations β, epistemic causality prescribes an algorithmic way of generating a causal theory $\mathcal{C}_\beta$. This is a different inductive approach to the causal-Markov methods most widely advocated today, but as I have argued in §3.2, those methods are based on questionable assumptions, and (§4.2) the epistemic causality approach explains the special cases where causal-Markov methods work.

On the other hand epistemic causality is hypothetico-deductive: a causal theory $\mathcal{C}_\beta$ is at best a tentative hypothesis, a set of beliefs, and needs testing before it can become entrenched as causal knowledge. Moreover epistemic causality provides a way of filling in the gaps of a hypothetico-deductive approach. The hypothetico phase is no mystery — we have an account of how a hypothesis $\mathcal{C}_\beta$ can be determined by knowledge of the indicators of causality.[58] The deductive phase is no mystery either: we test a causal hypothesis by the inverse mapping from causality to indicators. ¿From a causal relation we can predict a strategic dependency, the existence of a physical mechanism, a temporal relation, and so on, and the causal hypothesis is confirmed to the extent that those predictions are borne out.

5 PEARL'S DETERMINISM

In this section I shall compare epistemic causality with the position recently advocated by Judea Pearl, a pioneer of one of the inductive approaches for discovering causal relationships discussed in §3.2.

It is important to note that Pearl's recent views (as of 2000) differ significantly from his original conception of causality (of 1988).

Pearl's original position stressed the convenience of causality and had much in common with epistemic causality:[59]

> We take the position that human obsession with causation, like many other psychological compulsions, is computationally motivated. Causal models are attractive mainly because they provide effective data structures for representing empirical knowledge — they can be queried and updated at high speed with minimal external supervision.[60]

However, Pearl then changed his mind about causality altogether:

> Ten years ago, when I began working on *Probabilistic Reasoning in Intelligent Systems* (1988), I was working within the empiricist tradition. In this tradition, probabilistic relationships constitute the foundations of human knowledge, whereas causality simply provides useful ways of abbreviating and organizing intricate patterns of probabilistic relationships. Today, my view is quite different. I now take causal relationships to be the fundamental building blocks both of physical reality and of human

[58] Machine learning techniques can be used here to automate the generation of a hypothesis from a database of observations in conjunction with other background knowledge. See [Stankovski *et al.*, 2001] for an analogous proposal.

[59] Epistemic causality is compared to Pearl's early views in [Williamson, 2004], §9.4.

[60] [Pearl, 1988] 383.

> understanding of that reality, and I regard probabilistic relationships as but the surface phenomena of the causal machinery that underlies and propels our understanding of the world.[61]

Thus Pearl's new view is that causality is mind-independent and physical, not to be understood in terms of convenience of belief after all:

> . . . causal relationships are more "stable" than probabilistic relationships. We expect such difference in stability because causal relationships are *ontological*, describing objective physical constraints in our world, whereas probabilistic relationships are *epistemic*, reflecting what we know or believe about the world. Therefore, causal relationships should remain unaltered as long as no change has taken place in the environment, even when our knowledge about the environment undergoes changes.[62]

Interestingly, here Pearl appears to be invoking a physical notion of cause in order to account for the objectivity of causality. As I have pointed out in §4.3, this move is by no means necessary — equally one can account for objectivity by taking an epistemic approach. While for epistemic causality causal beliefs may change as knowledge changes, the induced fully objective notion of cause is independent of any particular agent's knowledge.

Pearl's recent view is that causal models are structural equation models (introduced in §3.2). Pearl's new account thus not only embraces physical causality, but also universal determinism:

> causal relationships are expressed in the form of deterministic, *functional* equations, and probabilities are introduced through the assumption that certain variables in the equations are unobserved. This reflects Laplace's (1814) conception of natural phenomena, according to which nature's laws are deterministic and randomness surfaces owing merely to our ignorance of the underlying boundary conditions.[63]

Pearl subsequently describes his reasons for preferring a deterministic approach to his more stochastic 1988 approach which took causal models to be causal nets rather than structural equation models:[64]

> First, the Laplacian conception is more general. Every stochastic model can be emulated by many functional relationships (with stochastic inputs), but not the other way around; functional relationships can only be approximated, as a limiting case,

[61] [Pearl, 2000] xiii-xiv.
[62] [Pearl, 2000] 25.
[63] [Pearl, 2000] 26.
[64] See also [Pearl, 2000], 31.

> using stochastic models. Second, the Laplacian conception is more in tune with human intuition. The few esoteric quantum mechanical experiments that conflict with the predictions of the Laplacian conception evoke surprise and disbelief, and they demand that physicists give up deeply entrenched intuitions about locality and causality. Our objective is to preserve, explicate, and satisfy — not destroy — those intuitions.
>
> Finally, certain concepts that are ubiquitous in human discourse can be defined only in the Laplacian framework. We shall see, for example, that such simple concepts as "the probability that event B occurred *because* of event A" and "the probability that event B would have been *different* if it were not for event A" cannot be defined in terms of purely stochastic models. These so-called *counterfactual* concepts will require a synthesis of the deterministic and probabilistic components embodied in the Laplacian model.[65]

While functional models may be desirable and appropriate in many circumstances, it seems perverse to develop a theory of causality that is inconsistent with indeterminism when indeterminism is advocated by our best scientific theories. Far better, in my view, to develop an account of causality that is consistent with indeterminism but to use deterministic functional models where possible. This is one of the advantages of epistemic causality over Pearl's later position: it leaves open the choice of model. According to epistemic causality, an agent's causal belief graph is purely qualitative, involving neither probabilistic relationships nor deterministic functional relationships. But this does not stop one from quantifying the causally connections using either type of relationship if it is appropriate to do so. Clearly an account that does not restrict one to appealing to just probabilistic relationships or to just deterministic relationships (i) is more general than either the purely stochastic or the purely deterministic approach, (ii) satisfies the demands of science as well as intuition, and (iii) can support Pearl's semantics for counterfactuals wherever deterministic models are appropriate.

Pearl's advocacy of the Causal Markov Condition is another point that sets it apart from epistemic causality. Because Pearl uses only structural equation models and assumes that the error variables are probabilistically independent, the Causal Markov Condition follows.[66] There are three difficulties with this justification. First it depends on the acceptance of universal determinism which, as we have seen, is problematic. Second, no independent argument is given for the assumption that error variables are independent. Pearl merely points out the utility of this assumption: it yields the Causal

[65] [Pearl, 2000] 26-27.
[66] [Pearl, 2000] Theorem 1.4.1.

Markov Condition and thereby agrees with the Principle of the Common Cause and the properties that ensue.[67]

Third, there are the counterexamples to the Causal Markov Condition referred to in §2.2. Pearl attempts to salvage the condition by arguing that counterexamples either belong to quantum mechanics (in which case they are ignorable for practical purposes) or they can be explained away by invoking latent variables (dummy variables that act as common causes).[68] However, the first response is undesirable both because the quantum domain is becoming increasingly important for technology (there is already considerable interest in applications of quantum computation and quantum cryptography), and because as yet it is just a matter of conjecture that quantum indeterminacy fails to infect the macroscopic world. The second response fails because while introducing latent variables can salvage the independencies posited by the Causal Markov Condition, the condition itself often still fails since it is often the case that a *causal interpretation* of a latent variable remains implausible (analogously if A and B are probabilistically dependent but neither causes the other, then the Principle of the Common Cause requires both that there be variables that render A and B independent, *and that these variables are interpretable as common causes* of A and B, not just dummy variables).[69] Pearl also claims that the continuing interest in probabilistic analyses of causality, which often invoke the Causal Markov Condition or an equivalent, lends weight to the condition: 'The intellectual survival of probabilistic causality as an active philosophical program for the past 30 years attests to the fact that counterexamples to the Markov condition are relatively rare and can be explained away through latent variables.'[70] This is rather flimsy evidence though: the history of philosophy is littered with failed attempts (lasting longer than 30 years) to produce a viable version of an initially attractive analysis.

Epistemic causality takes a different view. It accepts that counterexamples to the Causal Markov Condition do arise, but as we saw in §4.2, the condition demonstrably holds in certain special cases. This justifies a qualified use of Pearl's methods for causal reasoning and causal discovery (but not his ontology).

I have argued, then, that Pearl need not have changed his mind about the nature of causality in order to produce an objective notion of cause: epistemic causality, which does yield objectivity, can be viewed as close to Pearl's early approach. Moreover the unqualified adoption of deterministic causal models and the Causal Markov Condition leads to a formalism that is at best a first approximation to the complexity of causality. Epistemic causality aims to capture that complexity.

[67] [Pearl, 2000] 61.
[68] [Pearl, 2000] 62.
[69] [Williamson, 2004] §4.2.
[70] [Pearl, 2000] 62-63.

6 PRICE'S PRAGMATISM

Huw Price, a proponent of the agency theory discussed in §2.4, has developed an interesting 'perspectival' conception of causality that is based on pragmatism.

While pragmatism is normally associated with Peirce's and James' attempts to analyse truth in terms of belief (alluded to in §4.3), Price delineates his pragmatism as follows:[71]

> A third form of pragmatism, and the one that interests me here, is the view that a philosophical account of a problematic notion — that of causation itself, for example — needs to begin by playing close attention to the role of the concept concerned in the *practice* of the creatures who use it. Indeed, the need to explain the use of a notion in the lives of ordinary speakers is often the original motivation for an account of this kind. Causal notions and their kin are ubiquitous in the everyday talk of ordinary people. Pragmatists argue that we cannot hope to explain this anthropological fact if we begin where metaphysics traditionally begins, at the level of the objects themselves — if we ask what causation *is*, if we begin by looking for something for causation to *be*, which will explain all these uses. Instead, pragmatists think, we need to start with the practise of *using* such notions, and to ask what role such notions play in the lives of the creatures concerned — why creatures like us should have come to describe the world in these causal terms.[72]

The last sentence portrays pragmatism as the rather uncontroversial methodological claim that philosophical investigation of a problematic notion should start with an investigation of its use. Indeed epistemic causality takes practice (the convenience of causal representations) as a starting point and only then develops a more formal account of causality and of what causality is. However, there is more to Price's pragmatic account of causality than this advice as to where to begin. Price maintains that not only should one not *start* by asking what causality is, one should not ask what causality is at all — this is the wrong question and one should instead focus on how causal notions are *used*. (Epistemic causality, in contrast, makes no such claim; indeed it provides an account of what causality is.) On the other hand Price does narrow down what causality is. For Price causality is *perspectival*: causal models are viewed from an agent's standpoint,[73] but are projected onto the world,[74] and like fictions the perspectival aspect may not

[71] [Price, 2003] describes the relationship between his form of pragmatism and truth.
[72] [Price, 2001] 105.
[73] [Price, 2004] §3.1.
[74] [Price, 2004] §3.2.

be obvious to the agent:[75]

> Perhaps causal asymmetry isn't really in the world at all, but the *appearance that it is* is a product of our own standpoint. Perhaps it is like the warmth that we see when we look at the world through rose-tinted spectacles.[76]

Yet Price's notion of causality is not mental:

> let me emphasise that pragmatism about causation is not the view that when we talk of causation we are talking *about* ourselves, in whole or in part.[77]

> I simply want to emphasise that the view is not ... that talk of causation is talk *about* agents or agency, but rather the ... doctrine that we don't understand the notion of causation — as philosophers, as it were — until we understand its origins in the lives and practice of agents such as ourselves.[78]

This is another point of difference between Price's pragmatism and epistemic causality. Epistemic causality *is* a mental notion, in the sense that talk about causality is talk about what agents ought to believe. Since Price's conception of causality is not mental, his view is not analogous to the Bayesian view that probability is rational degree of belief.[79] In contrast, epistemic causality *is* analogous to this view: just as an agent ought to adopt a certain probability function as a representation of her degrees of belief, she ought to adopt a certain directed acyclic graph as a representation of her causal beliefs.[80] Moreover just as David Lewis viewed fully objective probabilities as those degrees of belief an agent ought to adopt were she to know everything relevant,[81] so too epistemic causality views a fully objective notion of cause as those causal beliefs an agent ought to adopt were she to know everything relevant.[82]

Note though that epistemic causality does not imply that if there were no agents there would be no causation — for epistemic causality causal beliefs are idealised, the beliefs that an agent ought to adopt, which remain well-defined in the absence of agents. Price concurs on this point:

> If the concept of causation is essentially tied to our experience as agents, as my kind of ... pragmatism maintains, then of course

[75] [Price, 2004] §3.3.
[76] [Price, 1996] 153.
[77] [Price, 2001] 107.
[78] [Price, 2001] 107.
[79] [Price, 2001] 107.
[80] [Williamson, 2004] §9.10.
[81] [Lewis, 1980].
[82] [Williamson, 2004] §9.9.

> the concept would not arise in a world without agents. But this does not make it appropriate to say that if there had been no agents there would have been no causation. Pragmatism does not conflict with realism in that sense.[83]

On the other hand Price goes on to argue that only an extremely weak form of realism remains tenable:

> This view simply takes the existence claims of science at face value, and rejects any 'additional' metaphysical or philosophical viewpoint from which it would really make sense to ask 'Do these things (electrons, for example) *really* exist?' The key to weak realism is a rejection of a standpoint for ontology beyond that of science.[84]

As Price acknowledges this is not much of a realist position:

> I am following convention in calling this view a species of realism. However, it is also instructive to see the view as rejecting the traditional realist-antirealist debate altogether, at least as that debate arises within the empiricist tradition.[85]

Epistemic causality is less radical. For epistemic causality the question of whether causal relations exist in the physical world does make sense; different varieties of epistemic causality (agnosticism and anti-physicalism) give different answers to this question.

Price advocates his 'weak realism' on the basis of the following problem with the more usual 'strong realism':

> the main argument for strong realism about theoretical entities goes in terms of inference to explanatory *causes*. But this reason simply takes the notion of causation for granted, and therefore can't be applied in *support* of realism about causation. In this context, the supposed role of inference to the best explanation is epistemological — it is supposed to *justify* a belief in the reality of entities of a certain kind. My point is that such an attempt at justification would be viciously circular in the case of causation itself, in virtue of the fact by the realist's own lights, the inference presupposes realism about explanatory causes.[86]

Note though that while Price does identify a potential problem for the view that causality is a physical relation, a dismissal of strong realism leaves

[83] [Price, 2001] 108.
[84] [Price, 2001] 112.
[85] [Price, 2001] 112.
[86] [Price, 2001] 113-114.

several tenable views — Price's own weak realism (a rejection of the realism-antirealism question) but also the anti-physical and agnostic varieties of epistemic causality — none of which appeal to inference to the best causal explanation. So Price's argument does not on its own decide between weak realism and epistemic causality.

A rather counter-intuitive relativity of the agency notion of causality might provide one deciding factor:

> Suppose that the world had developed in such a way that we had fewer manipulative abilities and skills than we actually possess but that we still applied our concept of causation roughly in conformity with the agency approach. In this case the reference of the expression 'relation between events such that an actual agent could manipulate one event as a means to bringing about the other' would have been fixed on different relations, even though our way of fixing the reference would have been the same.[87]

Thus the agency theory possesses a form of subjectivity: agents with different capacities may rationally disagree about causal relationships. This looks to be a problem not just across possible worlds but across agents in this world. Just as the capacities of a human, a robot and a Venus fly trap differ, so too would causality-for-a-human, causality-for-a-robot and causality-for-a-Venus-fly-trap. Such subjectivity is attributable to Price's view of causality as a secondary quality, like colour:

> we shall take as our reference point a simple version of the orthodox dispositional theory, namely the view that to be red is to be disposed to look red to a normal observer under standard conditions. This embodies the insight that colour is a secondary quality, defining the colour concept in terms of human capacities and responses. ... Our claim is simply that the agency theory correctly portrays causation as something analogous to a secondary quality — *as* a secondary quality, in fact, on a suitably extended understanding of that notion.[88]

However, while the subjectivity of colour does not clash strongly with intuition, causality does intuitively seem to be objective. Menzies and Price reply to this objection as follows:

> Our response is to accept that this kind of relativity is a consequence of the theories concerned, but to deny that it is untoward. We make two main points in support of this conclusion.

[87] [Menzies and Price, 1993] 199.
[88] [Menzies and Price, 1993] 188-189.

> The first, as usual, is that the characteristic of causation thus identified is already a non-problematic feature of colour and the other classical secondary qualities. It is something we live with in those cases, and may be expected to accommodate ourselves to in the case of causation. Secondly, however, we want to point out that there is an important difference of degree between the two cases. As we shall explain in a moment, it turns out that causality is very much less sensitive than colour, say, to the accidents of the human situation. In this we find a basis for the intuition that causation is significantly more 'objective' than the usual secondary qualities — an intuition with which we thus concur.[89]

Although the subjectivity of the agency theory of causality may be more limited than that of the dispositional theory of colour, and although some philosophers may be able to bite the bullet and live with the subjectivity, one can avoid the subjectivity altogether. Epistemic causality does not define causality in terms of agents' capacities and is not subjective in this problematic respect. Thus the objectivity of causality provides a reason to prefer epistemic causality over the agency account.

In sum, Price's objection to strong realism about causality need not force one to adopt his rather radical rejection of the realism-antirealism debate. Epistemic causality, which views causality as mental rather than physical, remains a contender. Moreover epistemic causality might be preferred over Price's agency theory, since the latter notion of causality suffers from relativity to the capacity of agents.

7 CONCLUDING REMARKS

We have seen that contemporary theories tend to explain causality in terms of just one of its indicators, in particular physical mechanisms, probabilistic relationships, functional relationships, counterfactual relationships or agency considerations. These approaches then find it hard to explain how all the other indicators can have a bearing on our causal judgements. However, by looking first at causal beliefs and the ways in which they are constrained by knowledge of these indicators, one can account for the complexity of causality. Moreover the ensuing approach, epistemic causality, provides an account of the objectivity of causality and an answer to fundamental questions about what causality is and how we can discover causal relationships.

There are a couple of philosophical concerns one might have with epistemic causality, to do with circularity.

[89] [Menzies and Price, 1993] 199-200.

The first concern is that the characterisation of epistemic causality might be circular. Epistemic causality provides an ultimate belief interpretation of a fully objective notion of cause. Thus causality is characterised in terms of causal beliefs. But if causal beliefs are beliefs about causality then the relationship between causality and causal beliefs is circular.

While this argument is valid, it does not tell against epistemic causality, for two reasons. First and foremost, epistemic causality provides an independent route to causal beliefs: in §4.1 causal beliefs are characterised independently of ultimate belief causality, in terms of knowledge of strategic dependencies, mechanisms, temporal relations, and so on.[90] Second, epistemic causality does not claim that causal beliefs are beliefs about causality. For epistemic causality, causal beliefs are a *type* of belief, not necessarily beliefs *about* anything in particular: 'causal' modifies 'beliefs' and does not specify an object of the beliefs. The claim that causal beliefs are beliefs about ultimate belief causality is in any case implausible: it is simply implausible to suggest that when Audrey believes that smoking causes cancer, she believes that were she to know about all the relevant indicators she ought to believe that smoking causes cancer. This latter point is perhaps more obvious when made regarding the Bayesian view of probability that is analogous to epistemic causality. Here the terminology 'degrees of belief' is used for 'probabilistic beliefs' while 'chance' is used for 'probability': degrees of belief are a type of belief and are not beliefs about chances. If they were beliefs about chances, then an ultimate belief characterisation of chance in terms of degrees of belief (such as that of Lewis) would be circular. But in any case it is implausible to suggest that when Bill believes that England will win the cricket to degree 0.8, he believes that were he to know the entire history of the world and all history-to-chance conditionals he would believe that England will win the cricket to degree 0.8.

The second worry is that the relationship between epistemic causality and its indicators might be circular. According to epistemic causality, causal beliefs depend on knowledge of the multifarious indicators of causality. If these indicators are themselves reducible to causal notions then it is natural to suspect circularity. For example, we might want to understand temporal direction in terms of causality — but how can this be possible if temporal knowledge helps delimit the causal relation? In contrast, if we simply reduce causality to counterfactuals then an account of temporal direction in terms of causality is more obviously non-circular.

In fact though, epistemic causality leaves open the question of which

[90] [Williamson, 2004] §9.8 deals with the case in which positive causal knowledge can constrain causal beliefs. In that case causal beliefs can depend upon ultimate belief causality. But there is no circularity there either, because ultimate belief causality is characterised in terms of causal beliefs relative to background knowledge that includes all knowledge of strategic dependencies, mechanisms and so on, but that does not include knowledge of ultimate causal relations.

reductive relationships obtain amongst its indicators. Epistemic causality offers a functional explanation of causality in terms of its convenience, and a characterisation of the causal relation in terms of rational beliefs, but not a reductive analysis of causality in terms of its indicators. Consider an analogy in medicine. When a condition is poorly understood, one may posit a *syndrome* and characterise it in terms of its indicators. For example, Tourette's syndrome is characterised (implicitly defined) in terms of involuntary tics and uncontrollable verbalisation, in particular the use of obscene language and the tendency to repeat uttered words. No commitment is made as to what actually causes what — indeed the causal picture regarding Tourette's syndrome is still unclear. As long as the characterisation of the syndrome latches onto something objective, it will suffice for diagnosis and treatment. Similarly, a characterisation of causality that latches onto something objective can offer a way of handling causality without presupposing relationships amongst its indicators: temporal direction can be a good indicator of causal direction whether or not the former is reducible to the latter.[91]

Thus epistemic causality offers a powerful alternative to the standard accounts of causality, yet one that is compatible with a range of philosophical agendas.[92]

Jon Williamson
University of Kent

BIBLIOGRAPHY

[Bacon, 1620] Francis Bacon: 'The New Organon', Lisa Jardine and Michael Silverthorne (eds.), Cambridge: Cambridge University Press 2000.

[Benacerraf, 1973] Paul Benacerraf: 'Mathematical truth', in [Benacerraf and Putnam, 1983], pages 403-420.

[Benacerraf and Putnam, 1983] Paul Benacerraf and Hilary Putnam (eds.): 'Philosophy of mathematics: selected readings', Cambridge: Cambridge University Press, second edition.

[Cartwright, 1997] Nancy Cartwright: 'What is a causal structure?', in [McKim and Turner, 1997], pages 343-357.

[Cartwright, 1999] Nancy Cartwright: 'Causality: independence and determinism', in [Gammerman, 1999], pages 51-63.

[Cartwright, 2001] Nancy Cartwright: 'What is wrong with Bayes nets?', The Monist 84(2), pages 242-264.

[91] Of course for the the anti-physicalist form of epistemic causality, any reduction of temporal direction to causal direction will yield a *mental* notion of temporal direction. In contrast, the agnostic form of epistemic causality leaves this issue open too: the mental epistemic account is one interpretation of causality, but there may be other viable interpretations of causality to which time (or other indicators) can be reduced.

[92] I am very grateful to Donald Gillies, Huw Price and Federica Russo for comments, and to Oxford University Press for permission to reprint passages from [Williamson, 2004] in §2 and §3.

[Cooper, 1999] Gregory F. Cooper: 'An overview of the representation and discovery of causal relationships using Bayesian networks', in [Glymour and Cooper, 1999], pages 3-62.

[Cooper, 2000] Gregory F. Cooper: 'A Bayesian method for causal modeling and discovery under selection', in Proceedings of the Conference on Uncertainty in Artificial Intelligence 2000, pages 98-106.

[Corfield and Williamson, 2001] David Corfield and Jon Williamson (eds.): 'Foundations of Bayesianism', Kluwer Applied Logic Series, Dordrecht: Kluwer Academic Publishers.

[Dai *et al.*, 1997] Honghua Dai, Kevin Korb, Chris Wallace and Xindong Wu: 'A study of causal discovery with weak links and small samples', in Proceedings of the 15th International Joint Conference on Artificial Intelligence (IJCAI-97), Nagoya, Japan, August 23-29 1997.

[Dash and Druzdzel, 1999] Denver Dash and Marek Druzdzel: 'A Fundamental Inconsistency Between Causal Discovery and Causal Reasoning', in Proceedings of the Joint Workshop on Conditional Independence Structures and the Workshop on Causal Interpretation of Graphical Models, The Fields Institute for Research in Mathematical Sciences, Toronto, Canada.

[Dawid, 2001] A.P. Dawid: 'Causal inference without counterfactuals', in [Corfield and Williamson, 2001], pages 37-74.

[Dowe, 1993] Phil Dowe: 'On the reduction of process causality to statistical relations', British Journal for the Philosophy of Science 44, pages 325-327.

[Dowe, 1996] Phil Dowe: 'Backwards causation and the direction of causal processes', Mind 105, pages 227-248.

[Dowe, 1999] Phil Dowe: 'The conserved quantity theory of causation and chance raising', Philosophy of Science 66 (Proceedings), pages S486-S501.

[Dowe, 2000] Phil Dowe: 'Causality and explanation: Review of Salmon', British Journal for the Philosophy of Science 51, pages 165-174.

[Dowe, 2000b] Phil Dowe: 'Physical causation', Cambridge: Cambridge University Press.

[Earman, 1992] John Earman: 'Bayes or bust?', Cambridge, Massachusetts: M.I.T. Press.

[Freedman and Humphreys, 1999] David Freedman and Paul Humphreys: 'Are there algorithms that discover causal structure?', Synthese 121, pages 29-54.

[Gammerman, 1999] Alex Gammerman (ed.): 'Causal models and intelligent data management', Berlin: Springer.

[Glymour, 1997] Clark Glymour: 'A review of recent work on the foundations of causal inference', [McKim and Turner, 1997], pages 201-248.

[Glymour, 2001] Clark Glymour: 'The Mind's Arrows: Bayes nets and graphical causal models in psychology', Cambridge, Massachusetts: The M.I.T. Press.

[Glymour and Cooper, 1999] Clark Glymour and Gregory F. Cooper (eds.): 'Computation, causation, and discovery', Cambridge, Massachusetts: The M.I.T. Press.

[Hagmayer and Waldmann, 2002] York Hagmayer and Michael R. Waldmann: 'A constraint satisfaction model of causal learning and reasoning', in Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society, Mahwah, NJ: Erlbaum.

[Hausman, 1999] Daniel M. Hausman: 'The mathematical theory of causation', review of [McKim and Turner, 1997], British Journal for the Philosophy of Science 50, pages 151-162.

[Hausman and Woodward, 1999] Daniel M. Hausman and James Woodward: 'Independence, invariance and the causal Markov condition', British Journal for the Philosophy of Science 50, pages 521-583.

[Heckerman *et al.*, 1999] David Heckerman, Christopher Meek and Gregory Cooper: 'A Bayesian approach to causal discovery', in [Glymour and Cooper, 1999], pages 141-165.

[Hempel and Oppenheim, 1948] Carl G. Hempel and Paul Oppenheim: 'Studies in the logic of explanation', with Postscript in [Pitt, 1988], pages 9-50.
[Howson and Urbach, 1989] Colin Howson and Peter Urbach: 'Scientific reasoning: the Bayesian approach', Chicago: Open Court, Second edition, 1993.
[Humphreys, 1997] Paul Humphreys: 'A critical appraisal of causal discovery algorithms', in [McKim and Turner, 1997], pages 249-263.
[Humphreys and Freedman, 1996] Paul Humphreys and David Freedman: 'The grand leap', British Journal for the Philosophy of Science 47, pages 113-123.
[James, 1907] William James: 'What pragmatism means', in 'Pragmatism: A new name for some old ways of thinking', New York: Longman Green and Co, pages 17-32.
[Karimi and Hamilton, 2000] Kamran Karimi and Howard J. Hamilton: 'Finding Temporal Relations: Causal Bayesian Networks versus C4.5', Proceedings of the Twelfth International Symposium on Methodologies for Intelligent System (ISMIS'2000), Charlotte, NC, USA, October 2000.
[Karimi and Hamilton, 2001] Kamran Karimi and Howard J. Hamilton: 'Learning causal rules', Technical Report CS-2001-03, Department of Computer Science, University of Regina, Saskatchewan, Canada.
[Korb, 1999] Kevin B. Korb: 'Probabilistic causal structure', in H. Sankey (ed.): 'Causation and laws of nature', Dordrecht: Kluwer, pages 265-311.
[Korb and Nicholson, 2003] Kevin B. Korb and Ann E. Nicholson: 'Bayesian artificial intelligence', London: Chapman and Hall / CRC Press UK.
[Lad, 1999] Frank Lad: 'Assessing the foundation for Bayesian networks: a challenge to the principles and the practice', Soft Computing 3(3), pages 174-180.
[Lemmer, 1996] John F. Lemmer: 'The causal Markov condition, fact or artifact?', SIGART Bulletin 7(3), pages 3-16.
[Lewis, 1973] David K. Lewis: 'Causation', with postscripts in [Lewis, 1986], pages 159-213.
[Lewis, 1980] David K. Lewis: 'A subjectivist's guide to objective chance', in [Lewis, 1986], pages 83-132.
[Lewis, 1986] David K. Lewis: 'Philosophical papers volume II', Oxford: Oxford University Press.
[Lewis, 1986b] David K. Lewis: 'Causal explanation', in [Lewis, 1986], pages 214-240.
[Lewis, 2000] David K. Lewis: 'Causation as influence', The Journal of Philosophy 97(4), pages 182-197.
[Mani and Cooper, 1999] Subramani Mani and Gregory F. Cooper: 'A study in causal discovery from population-based infant birth and death records', in Proceedings of the AMIA Annual Fall Symposium 1999, Philadelphia: Hanley and Belfus Publishers, pages 315-319.
[Mani and Cooper, 2000] Subramani Mani and Gregory F. Cooper: 'Causal discovery from medical textual data', in Proceedings of the AMIA annual fall symposium 2000, Philadelphia: Hanley and Belfus Publishers, pages 542-546.
[Mani and Cooper, 2001] Subramani Mani and Gregory F. Cooper: 'Simulation study of three related causal data mining algorithms', in Proceedings of the International Workshop on Artificial Intelligence and Statistics 2001, San Francisco: Morgan Kaufmann, pages 73-80.
[McKim and Turner, 1997] Vaughn R. McKim and Stephen Turner: 'Causality in crisis? Statistical methods and the search for causal knowledge in the social sciences', Notre Dame: University of Notre Dame Press.
[Menzies and Price, 1993] Peter Menzies and Huw Price: 'Causation as a secondary quality', British Journal for the Philosophy of Science 44, pages 187-203.
[Pearl, 1988] Judea Pearl: 'Probabilistic reasoning in intelligent systems: networks of plausible inference', San Mateo, California: Morgan Kaufmann.
[Pearl, 1999] Judea Pearl: 'Graphs, structural models, and causality', in [Glymour and Cooper, 1999], pages 95-138.

[Pearl, 2000] Judea Pearl: 'Causality: models, reasoning, and inference', Cambridge: Cambridge University Press.

[Peirce, 1877] Charles Sanders Peirce: 'The fixation of belief', Popular Science Monthly 12, pages 1-15.

[Pitt, 1988] Joseph C. Pitt (ed.): 'Theories of explanation', Oxford: Oxford University Press.

[Polya, 1945] George Polya: 'How to solve it', second edition, Penguin 1990.

[Polya, 1954] George Polya: 'Induction and analogy in mathematics', volume 1 of 'Mathematics and plausible reasoning', Princeton: Princeton University Press.

[Polya, 1954b] George Polya: 'Patterns of plausible inference', volume 2 of 'Mathematics and plausible reasoning', Princeton: Princeton University Press.

[Popper, 1934] Karl R. Popper: 'The Logic of Scientific Discovery', with new appendices of 1959, London: Routledge 1999.

[Price, 1991] Huw Price: 'Agency and probabilistic causality', British Journal for the Philosophy of Science 42, pages 157-176.

[Price, 1992] Huw Price: 'Agency and causal asymmetry', Mind 101, pages 501-520.

[Price, 1992b] Huw Price: 'The direction of causation: Ramsey's ultimate contingency', Philosophy of Science Association 1992(2), pages 253-267.

[Price, 1996] Huw Price: 'Time's arrow and Archimedes' point: new directions for the physics of time', New York: Oxford University Press.

[Price, 2001] Huw Price: 'Causation in the special sciences: the case for pragmatism', in Domenico Costantini, Maria Carla Galavotti and Patrick Suppes (eds.): 'Stochastic Causality', Stanford, California: CSLI Publications, pages 103-120.

[Price, 2003] Huw Price: 'Truth as convenient friction', Journal of Philosophy 100, pages 167-190.

[Price, 2004] Huw Price: 'Models and modals', in Donald Gillies (ed.): 'Laws and models in science', London: King's College Publications.

[Railton, 1978] Peter Railton: 'A deductive-nomological model of probabilistic explanation', in [Pitt, 1988], pages 119-135.

[Reichenbach, 1956] Hans Reichenbach: 'The direction of time', Berkeley and Los Angeles: University of California Press 1971.

[Russell, 1913] Bertrand Russell: 'On the notion of cause', Proceedings of the Aristotelian Society 13, pages 1-26.

[Salmon, 1980] Wesley C. Salmon: 'Causality: production and propagation', in [Sosa and Tooley, 1993], chapter 9.

[Salmon, 1980b] Wesley C. Salmon: 'Probabilistic causality', in [Salmon, 1998], pages 208-232.

[Salmon, 1984] Wesley C. Salmon: 'Scientific explanation and the causal structure of the world', Princeton: Princeton University Press.

[Salmon, 1997] Wesley C. Salmon: 'Causality and explanation: a reply to two critiques', Philosophy of Science 64(3), pages 461-477.

[Salmon, 1998] Wesley C. Salmon: 'Causality and explanation', Oxford: Oxford University Press.

[Scheines, 1997] Richard Scheines: 'An introduction to causal inference', in [McKim and Turner, 1997], pages 185-199.

[Sosa and Tooley, 1993] Ernest Sosa and Michael Tooley (eds.): 'Causation', Oxford: Oxford University Press.

[Spirtes *et al.*, 1993] Peter Spirtes, Clark Glymour and Richard Scheines: 'Causation, Prediction, and Search', Cambridge, Massachusetts: The M.I.T. Press, second edition 2000.

[Stankovski *et al.*, 2001] V. Stankovski, I. Bratko, J. Demsar and D. Smrke: 'Induction of hypotheses concerning hip arthoplasty: a modified methodology for medical research', Methods of Information in Medicine 40, pages 392-396.

[Suppes, 1970] Patrick Suppes: 'A probabilistic theory of causality', Amsterdam: North-Holland.

[Tenenbaum and Griffiths, 2001] Joshua B. Tenenbaum and Thomas L. Griffiths: 'Structure learning in human causal induction', in T. Leen, T. Dietterich, and V. Tresp (eds.): Advances in Neural Information Processing Systems 13, Cambridge, Massachusetts: The M.I.T. Press, pages 59-65.

[Tong and Koller, 2001] Simon Tong and Daphne Koller: 'Active Learning for Structure in Bayesian Networks', in B. Nebel (ed.): Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence, San Francisco: Morgan Kaufmann, pages 863-869.

[Waldmann, 2001] Michael R. Waldmann: 'Predictive versus diagnostic causal learning: Evidence from an overshadowing paradigm', Psychonomic Bulletin and Review 8, pages 600-608.

[Waldmann and Martignon, 1998] Michael R. Waldmann and Laura Martignon: 'A Bayesian network model of causal learning', in M.A. Gernsbacher and S.J. Derry (eds.): 'Proceedings of the Twentieth Annual Conference of the Cognitive Science Society', Mahwah, New Jersey: Erlbaum, pages 1102-1107.

[Wallace and Korb, 1999] Chris S. Wallace and Kevin B. Korb: 'Learning linear causal models by MML sampling', in [Gammerman, 1999], pages 88-111.

[Wendelken and Shastri, 2000] Carter Wendelken and Lokendra Shastri: 'Probabilistic Inference and Learning in a Connectionist Causal Network', In Proceedings of the Second International Symposium on Neural Computation, Berlin, May 2000.

[Williamson, 2004] Jon Williamson: 'Bayesian nets and causality: philosophical and computational foundations', Oxford: Clarendon Press.

[Woodward, 1997] James Woodward: 'Causal models, probabilities, and invariance', in [McKim and Turner, 1997], pages 265-315.

[Yoo *et al.*, 2002] Changwon Yoo, V. Thorsson and G. Cooper: 'Discovery of Causal Relationships in a Gene-regulation Pathway from a Mixture of Experimental and Observational DNA Microarray Data', in Proceedings of the Pacific Symposium on Biocomputing, New Jersey: World Scientific, pages 498-509.

DOROTHY EDGINGTON

ON CONDITIONALS*

The ability to think conditional thoughts is a basic part of our mental equipment. A view of the world would be an idle, ineffectual affair without them. There's not much point in recognising that there's a predator in your path unless you also realise that if you don't change direction pretty quickly you will be eaten.

Happily, we handle ifs with ease. Naturally, we sometimes misjudge them, and sometimes don't know what to think. But we know what it would take to be in a position to think or say that *B* if *A*, what would count for or against such judgements, how they affect what we should do and what else we should think. They cause us no undue practical difficulty.

The theory of this practice is another story. Judged by the quality and intensity of the work, theorising about conditionals has flourished in recent years — bold, fertile ideas developed with ingenuity and rigour, hitherto unnoticed phenomena observed and explained, surprising results proved. But consensus has not emerged. Not just about details, but about fundamentals, almost everything is at issue. Is a unified theory possible, or are there irreducibly different kinds of 'if'? If the latter, what marks the distinction between kinds, and which examples belong together? Is the core of a theory a thesis about what makes a conditional statement true? Those who suppose so dispute about the kind of truth conditions involved; others think it is a mistaken presumption that conditionals are part of fact-stating discourse, evaluable in terms of truth. Given these disputes, it is unsurprising that there are disagreements about which inference patterns involving conditionals are valid. There is even dissent about the logical form of conditionals: we are already theorising when we represent a conditional as a particular mode of combining two simpler propositions into one, and this representation has been questioned.

1 ONE THEORY OR TWO?

1.1

Something must be said at the outset about the classification of conditionals into kinds, for some theories address one kind, some another. Traditionally, 'indicative conditionals' have been distinguished from 'subjunctive conditionals' or 'counterfactuals' (these latter terms being used interchangeably). Some works concern conditionals of these forms:

(1a) If the gardener didn't do it, the butler did;

(1b) If the gardener doesn't do it, the butler will.

*The original version of this paper appeared in *Mind*, 104, 414, April 1995, pp. 235–329. It is reprinted here with the Editor's permission. Section 10 has been rewritten and Section 9.3 substantially revised. Elsewhere footnotes have been added concerning recent work.

D.M. Gabbay and F. Guenthner (eds.), Handbook of Philosophical Logic, 2nd Edition, Volume 14, 127–221.

For instance, W. V. Quine, in *Methods of Logic*, writes 'the contrafactual conditional is best dissociated from the ordinary conditional in the indicative mood... We shall not recur to it here' [1952, p. 21]. Other works concern those like

(1c) If the gardener had not done it, the butler would have;

(1d) If the gardener were not to do it, the butler would do it.[1]

For instance, David Lewis's *Counterfactuals*. 'I cannot claim to be giving a theory of conditionals in general', he says.

'There are different kinds of conditionals' can be taken as an innocuous remark, inevitably true. But the traditional distinction is less between two species of a genus, than between two genera, requiring separate treatment. This can surprise, for, it would seem, the sample sentences above could each be used to express the same conditional thought on different occasions. Changing the example: we are arguing about whether, if you eat this apple, you will be ill. You throw it away in disgust. Our argument continues unabated — about whether you would have been ill if you had eaten it. We do not appear to have changed the topic of debate. Just before throwing it away, you say 'If I were to eat it, ...'; someone who left our company earlier says later on 'I'm convinced that if he ate the apple, he was ill'. The bipartite approach needs some explanation.

Part of the explanation is, I think, historical. 'Contrary-to-fact' or 'subjunctive' conditionals first surfaced as a problem in the philosophy of science, for the attempt by logical empiricists to regiment scientific language using Frege's powerful new logic — to do for science what Frege and Russell had done for mathematics. At the heart of this logic is a treatment of the conditional of remarkable simplicity and clarity: a conditional is true if and only if it is not the case that it has a true antecedent and a false consequent.[2] When it came to analysing dispositional predicates like 'soluble' and 'fragile', a different kind of conditional made its presence felt. Being fragile is being such as to break if dropped. Frege's analysis cannot be used here; for it cannot explain why, if the vase is not dropped at a particular time, that doesn't settle whether the vase is fragile at that time — whether it would break if it were dropped. Rudolf Carnap's 'Testability and Meaning' [Carnap, 1936] was a valiant attempt to deal with this problem in terms of Fregean logic. Domesticating the non-Fregean conditional became a major problem.

If a theory fits some but not all of the data, the lesson might be that the data are not amenable to uniform treatment; it might equally be that we need a better theory. But there is more to be said in favour of dualism about conditionals, independently

[1] Throughout this section the letters a–d, following the numbers, indicate the form of the conditional.

[2] See Frege's *Begriffsschrift*, Section 5 [Frege, 1960, pp, 5–6]; his letter to Husserl translated in Frege [1980, p. 69]; and his 'Introduction to Logic' in Frege [1979, p. 186]. (In the latter two passages, which are comments on the first, Frege explains the conditional as I did above. In the *Begriffsschrift* he has 'affirmed' and 'denied' in place of 'true' and 'false' [Frege, 1960], p. 5. It is difficult to interpret this plausibly. On the following page he has 'to be affirmed' and 'to be denied', which, in the context, can more plausibly be interpreted as 'true' and 'false'.)

of prior theoretical commitment. Ernest Adams [1970] made the point with this striking pair of examples:

(2a) If Oswald didn't kill Kennedy, someone else did;

(2c) If Oswald hadn't killed Kennedy, someone else would have.

Everyone who knows of Kennedy's assassination agrees with (2a); many such people dissent from (2c). Take someone who thinks Oswald did it, acting alone. 'But what is the case if he didn't do it?' gets one answer. 'But what would have been the case if he hadn't done it?' gets a different answer. (Like any good philosophical example, this is no isolated case. Once you grasp its structure, you have a recipe for constructing indefinitely many such pairs, which I shall call 'OK cases'.) 'Therefore there really are two different sorts of conditional', says Lewis [1973, p. 3], commenting on this phenomenon, 'not a single conditional that can appear as indicative or as counterfactual depending on the speaker's opinion about the truth of the antecedent'. Here is how one might fill out the argument for this conclusion. Consider the two past-tense sentences:

O: Oswald didn't kill Kennedy

S: Someone else killed Kennedy

and consider the sentence frames:

If it is the case that..., it is the case that...

If it were the case that..., it would be the case that...[3]

Substitute the two sentences in the two sentence frames, and you have regimentations of (2a) and (2c). Replace each sentence frame by a symbol to be written between the sentences, say '$\rightarrow$', and '$\Box\!\!\rightarrow$', respectively. So we have '$O \rightarrow S$', '$O \Box\!\!\rightarrow S$'. One may accept '$O \rightarrow S$' yet reject '$O \Box\!\!\rightarrow S$'. So '$\rightarrow$' and '$\Box\!\!\rightarrow$' don't mean the same. QED.

This argument for two meanings of 'If...' is resistible. Our regimentations may have misrepresented the syntactic structure of the two sentences. Even when a single sentence has a true and a false reading, it does not follow that one of its semantic components is ambiguous. Consider

The Prime Minister has never been a woman.

That has a true and a false reading, but it is a case of syntactic, rather than semantic, ambiguity: the sentence may be read as structured in different ways, though each of its components has a uniform meaning. Or consider

[3]Lewis [1973, pp. 2–3], explains the counterfactual connective in terms of this second sentence frame.

> I *could have* been in New York today; but I *can't*, now, be in New York today.

This is not an example of an ambiguity in a modal term: something *was* possible, which *is no longer* possible.

This last example is instructive. V. H. Dudman[4] has convinced many that (2c): 'If Oswald hadn't killed Kennedy, someone else would have' is simply the past tense of

> (2b) If Oswald doesn't kill Kennedy, someone else will.

'Would have' is the past tense of 'will', as 'could have' is the past tense of 'can'; the verb forms in the antecedents typically indicate that they concern a time earlier than the consequents.[5]

The analysis of the counterfactual as a past-tense indicative could be a step in the direction of monism. If we can explain how the evaluation of a conditional depends on time, we can explain the OK cases without multiplying senses of 'if'. But this is not the moral drawn by Dudman and others. They remain dualists, and retain the view that (2a) and (2c) are different kinds, but maintain that (2b) has been wrongly classified: it is of a kind with (2c), not with (2a).[6]

The OK phenomenon does not support this new line, however, for it can be used to drive a wedge between future indicatives and counterfactuals as well as past ones. You think that such-and-such will happen. You can distinguish the questions: 'But what if it doesn't?' (i.e., what if you're wrong in thinking it will?); and 'But what if it were not going to?' (retaining your belief that it will). For instance, there are two prisoners, Smith and Jones. We have powerful evidence that one of them will try to escape tonight. Smith is a docile, unadventurous chap, Jones just the opposite, and very persistent. We are inclined to think that it is Jones who will try to escape. We have no reason to accept:

> (3c) If Jones were not to try to escape tonight, Smith would.

However, we could be wrong in thinking that it is Jones who will escape:

> (3b) If Jones doesn't try to escape tonight, Smith will.

Another example: I'm being chased through enemy territory, and a warning light on my (eccentric) car indicates that either I am about to run out of fuel, or the

[4] See e.g. Dudman [1983; 1984a; 1988; 1989]. Adams [1975], Ayers [1965] and Ellis [1984] also treat the 'counterfactual' as a past tense conditional.

[5] This is not how Dudman would put it. He does not care for the terms 'antecedent' and 'consequent' [Dudman, 1986; Dudman, 1988].

[6] See [Dudman, 1984], [Smiley, 1984], [Bennett, 1988], [Mellor, 1993].

radiator is about to boil over. I'm pretty sure it's the fuel. Bother! If I hadn't been going to run out of fuel, I would get away. Of course, I could be wrong about the fuel. But then, if I don't run out of fuel, the radiator will boil over.

The difference marked by the OK cases seems to be the traditional one. But it may be more like the difference between mature cheddar and freshly-made cheddar than the difference between chalk and cheese. As time passes but relevant information stays the same, 'If he eats the apple,...', 'If he were to eat it,...', 'If he ate it,...' and 'If he had eaten it,...' may all express the same conditional thought. But the passing of time may bring new relevant information: 'If he ate it, it did him no harm; but if he had eaten it, he would have been ill'. Further argument will have to wait on whether this difference can be explained within a unified account of 'if'.

1.2

The terminology for the traditional distinction is less than satisfactory. For those who accept the distinction, this is a minor irritant; for those who don't, it is a symptom of confusion. Lewis says 'You may justly complain that my title "Counterfactuals" is too narrow for my subject. I agree, but I know better.... The title "Subjunctive Conditionals" would not have delineated my subject properly [either]' [Lewis, 1973, pp. 3–4]. Long before, Roderick Chisholm announced that he would use 'subjunctive' and 'contrary-to-fact' interchangeably, although they were not coextensive. 'Neither term is adequate' [Chisholm, 1946, p. 482]. Michael Ayers complained that it was 'as if he had said that some mammals are not carnivores and some carnivores are not mammals, but he wished to talk of an important class of animal to which ... he would refer indiscriminately as mammals and as carnivores' [Ayers, 1965, p. 348]. Jonathan Bennett's 'Farewell to the Phlogiston Theory of Conditionals' [1988] also argues that the terminological inadequacy is a sign that our theories are in bad shape.

A true counterfactual may have a true antecedent and consequent, according to accepted usage. Consider 'If you had dropped it, it would have broken'. 'You're right — I did drop it, and it broke, but I did such a marvellous repair job, you never could tell'. Still, the idea behind the name is that counterfactuals are *for* talking about unrealised possibilities — we use them when we think the antecedent is false. But there is one important use of the 'counterfactual' form which does not fit this pattern. Alan Ross Anderson [1951] gave the example of a doctor's saying 'If he had taken arsenic, he would have shown just these symptoms [those which he in fact shows]'. The doctor could not convey the same thing with 'If he took arsenic, he is showing just these symptoms'. This is no one-off example: 'A bus is coming.' 'How do you know?' (for we can't see the oncoming traffic). 'People in line are picking up their bags and inching forward — *and that's what they would be doing if a bus were coming.*' (Not: that's what they are doing if a bus is coming.) These cases are important as ingredients in 'inference to the best explanation' and in Bayesian reasoning: which hypothesis H is such that what we do observe is

what we would expect to observe, if *H* were true?

(Conversely, it is sometimes the indicative which is needed to express disbelief in the antecedent: 'If he took arsenic, he's showing no signs'. Not: 'If he had taken arsenic, he would be showing no signs'.)

Would 'Subjunctive Conditionals' have been a better title for Lewis's book? Dudman [1988] and Bennett [1988] argue that the 'had been' and 'would' are a matter of tense, not mood. They quote grammarians who pour scorn on the idea that the subjunctive has any serious use in English. Grammarians are no more prone to unanimity than philosophers, however: *Fowler's Modern English Usage* gives the examples

> *If he heard, he gave no sign* (*heard* and *gave* past time); and *If he heard, how angry he would be!* (*heard* and *would be*, not past time, but utopia, the realm of non-fact or the imaginary); the first *heard* is indicative, the second subjunctive. [Fowler, 1965, p. 597].

Even if this is right, we lack a good explanation of why some conditionals require this mood and others forbid it. Further illumination is unlikely in advance of some theorising. It will cause less confusion and no greater offence if I stick to the labels 'indicative' and 'counterfactual' when discussing theories addressing one or the other side of the traditional divide.

2 TRUTH CONDITIONS OF THE FIRST KIND

2.1

There are conditional questions, commands, expressions of wish, etc., as well as conditional statements; but we follow the methodology of mainstream philosophy of language if we assume that an understanding of fact-stating discourse is our first task. Put counterfactuals aside. Assume that the conditional is a device for constructing a proposition, apt for truth, out of two component propositions, apt for truth. And it is a systematic device: if you understand any conditional, you understand every conditional whose components you understand. Still following the mainstream, assume that understanding a sentence is knowing under what circumstances it would be true. Understanding a sub-sentential meaningful component is knowing what contribution it makes to the truth conditions of the sentences in which it occurs. Some such components are used to construct complex sentences out of simpler sentences. Let *M* be: Mary went to Paris. Let *J* be: John went to Paris. Consider

(1a) It is not the case that *M*;
(1b) It is possible that, probable that, important that, relevant that *M*;
(2a) *M* and *J*, *M* or *J*;

(2b) *M* before *J*, *M* because *J*.

(l a) and (2a) are operators with a peculiarly simple property: in any possible circumstance, the truth value of the complex sentence is fixed by the truth value(s) of the simple sentence(s). Thus we write truth tables, showing the truth value of the whole for different possible combinations of the truth values of the parts; they are the truth-functional sentence operators. (1b) and (2b) lack this simple property. The truth values of the parts are not always sufficient to determine the truth value of the whole. They are non-truth-functional sentence operators. We need to examine the thesis that 'if' is truth-functional.

2.2

There are signs of it in ancient times, and it is sometimes called the Philonian conditional after Philo of Megara of the 4th century BC, but it is to Frege that we owe its role in current thinking about conditionals. It is a cornerstone of his system of logic, taken up enthusiastically by Russell (who called it 'material implication'), Wittgenstein and the logical positivists, and is now found in every logic book. It is the first theory of the conditional that students of philosophy encounter. And it has many defenders. We have already seen the one-liner: 'If *A* then *B*' is true if and only if it's not the case that *A* is true and *B* is false. It is thus equivalent to $\neg(A\&\neg B)$ and to $\neg A \vee B$. '$A \supset B$' has, by stipulation, these truth conditions. The substantive question is whether this is an adequate rendering of 'If A, B'.

It is easy to see that *if* 'if' is truth-functional, this is the right truth-function to assign it. For no one doubts[7] that a conditional is *sometimes* true when the truth values of its components are (true, true), or (false, true), or (false, false). Given truth-functionality, it follows that it is *always* true in these circumstances — for the truth-values of the components fix the truth value of the whole. Take a conditional which is true come what may, for example 'If Mary and John are both in Paris, then Mary is in Paris'. The components are such that it is impossible that it has a true antecedent and false consequent. But the other three combinations are possible, and whichever obtains, the conditional is true. Given truth-functionality, it follows that whenever one of these three combinations obtains, a conditional is true.

2.3

But is 'if' truth-functional? There are powerful arguments that it must be. No one denies that 'If A, B' entails $\neg(A\&\neg B)$, which is equivalent to $\neg A \vee B$. If the converse entailment holds, the truth-functional account is right. Getting the negation signs in more digestible places, the issue is equivalent to whether (i) $A \vee B$ entails 'If $\neg A, B$'; or (ii) $\neg(A\&B)$ entails 'If $A, \neg B$'.[8] But surely they do!

[7] No one who speaks of truth for conditionals at all, that is.

[8] (i) Let $A = \neg C$. Then $A \vee B$ entails 'If $\neg A, B$' iff $\neg C \vee B$ entails 'If $\neg\neg C, B$', i.e. 'If C, B' (given double negation elimination); (ii) Let $B = \neg D$; then $\neg(A\&B)$ entails 'If $A, \neg B$' iff $\neg(A\&\neg D)$ entails

Knowing just that at least one of the propositions, A, B, is true, is enough to infer that if A is not true, B is true; and (ii) knowing just that A and B are not both true is enough to infer that if A is true, B is not. For example: (i) having eliminated all but two suspects, I'm sure that either the gardener or the butler did it. So, if the gardener didn't do it, the butler did (water the aspidistra, that is); (ii) knowing that Mary and John were not both there, I infer that if Mary was there, John was not.

Putting the matter the other way round, suppose $A \vee B$ did not entail 'If $\neg A, B$' (but the propositions are compatible). Then these are two distinct possibilities:

	$A \vee B$	'If $\neg A, B$'
1.	T	T
2.	T	F

Suppose you are certain that one of these two possibilities obtains — but minimally so: you have eliminated $\neg A \& \neg B$, nothing more. This would not be enough for certainty that if $\neg A$, B because the possibility at line 2 would be compatible with your information. But, we have seen above, minimal certainty that $A \vee B$ is enough for certainty that if $\neg A$, B. Only the truth-functional truth conditions get this right: any stronger truth conditions get this wrong.

Again, here is a little proof of one of the crucial entailments. We make three assumptions: (i) $\neg(A \& B)$; (ii) A; (iii) B. We derive a contradiction. So, keeping assumptions (i) and (ii), we derive $\neg B$. So, by Conditional Proof, keeping assumption (i), we derive 'If A, $\neg B$'.

2.4

So what's the snag? Well, it seems strange to say that the falsity of 'She ate the apple', is sufficient for the truth of 'If she ate the apple, she was ill', as it is on this account. ($\neg A$ entails $\neg(A \& \neg B)$ for any B; let B be: she was ill). And this kind of example is the source of a catalogue of oddities. But perhaps it seems strange for the following reason. When we consult our intuitions about the inference from 'She didn't eat the apple', we imagine ourselves certain of that premiss. Then we don't have any serious use for a conditional that begins 'If she ate the apple'. If a theory which serves us well most of the time has the consequence that all such uninteresting conditionals are true, perhaps we can and should live with that consequence. It is too much — or maybe too little — to expect our theories to match ordinary usage perfectly. Perhaps, in the interests of simplicity and clarity, we should replace 'if' with '$\supset$'.

We should not. The unacceptability of the inference from $\neg A$ to 'If A, B' emerges most clearly in the context of beliefs which are less than certain. The problem was invisible to Frege and Russell (among many others): their main tar-

'If $A, \neg\neg D$', i.e. 'If A, D'. (Intuitionist worries about double negation elimination can be waived by assuming that the propositions are decidable.)

get was mathematical reasoning; holding beliefs on less-than-certain grounds was not in their main line of business. The worst defects of the truth-functional conditional don't show up in mathematics.

I shall use 'think that', 'believe' and 'disbelieve' in such a way as not to imply certainty. If you believe P, and disbelieve C, and there is a simple, decidable, valid argument from P to C, your beliefs are irrational. I have in mind things like: believing that something is square but disbelieving that it has 4 sides; believing that John and Mary are in Paris but disbelieving that John is in Paris. If P entails C, there is no way that P can be true without C being true. If the entailment is obvious, you should not be more confident that P is true than you are that C is true.

When I think, but am not certain, that $\neg P$, it is not at all uninteresting or unimportant to contemplate what is true if P. For example, (i) I think that my husband isn't home yet. But if he is, he'll be worried about where I am. So I should try to phone. Compare (ii): I think that the Queen isn't home yet (at Buckingham Palace, that is). But if she is, she'll be worrying about where I am. So I should try to phone. The first thoughts are sane enough, the second a sign of madness. Not so on the truth-functional account. Suppose, having read in the newspaper of her day's engagements, I'm about 90% certain that the Queen isn't at home yet ($\neg Q$); then I must be at least 90% certain that at least one of the propositions $\{\neg Q, W\}$ is true, i.e. at least 90% certain that $\neg Q \vee W$, i.e. at least 90% certain that if she is at home, she is worrying about my whereabouts (on the truth-functional reading of that thought). Someone who believes $\neg Q$, but disbelieves 'If Q,W' (on this reading) is making an Incredibly Gross Logical Error. For to disbelieve $Q \supset W$, i.e. $\neg(Q \& \neg W)$, is to believe its negation, $Q \& \neg W$. How can anyone be so stupid as to believe $Q \& \neg W$ yet disbelieve Q, i.e. believe $\neg Q$?

Contrary to this account, any sane ordinary subject not on intimate terms with royalty, who thinks the Queen isn't home yet, rejects the conditional 'But if she is, she'll be worried about where I am'. We do not use conditionals as this account would have it. But that empirical observation is not the main point, which is this: we would be intellectually disabled without the ability to discriminate between believable and unbelievable conditionals whose antecedents we think are unlikely to be true. The truth-functional account deprives us of this ability: to judge A unlikely is to commit oneself to the probable truth of $A \supset B$.

2.5

In his William James lectures, 'Logic and Conversation', delivered in 1967, H. P. Grice defended the truth-functional account, emphasising the importance of distinguishing the false from the misleading-but-true (see [Grice, 1989]). There are many ways of speaking the truth yet misleading your audience, given the standards to which you are expected to conform in conversation. One way is to say something weaker than some other relevant thing you are in a position to say. Consider disjunctions. I am asked where John is. I'm sure he's in the bar, and I know he

never goes near libraries. Inclined to be unhelpful but not wishing to lie, I say 'He's either in the bar or in the library'. My hearer naturally concludes that this is the most precise information I am in a position to give, and also concludes from the truth (let us assume) that I told him 'If he's not in the bar he's in the library'. The conditional, like the disjunction, according to Grice, is true provided he is in the bar, but misleadingly asserted on that ground.

Again: 'You won't eat those and live', I say of some wholesome and delicious mushrooms — knowing that you will now leave them alone, deferring to my expertise. I told no lie — for indeed you don't eat them — but of course I misled you. (Lewis [1976, p. 143] uses this example.)

Grice drew attention, then, to situations in which a person is justified in believing a proposition, which would nevertheless be an unreasonable thing for the person to *say*, in normal circumstances. His lesson was salutary in many areas of philosophy: the oddity of remarking under normal conditions of observation (e.g.) 'The pillar-box seems red to me', does not show that that sentence is false, or truth-value-less, or meaningless in that context.[9] The remark is true, but misleading unless you have a reason for doubting that it *is* red. Grice also explains correctly the behaviour of disjunctions and negated conjunctions. Believing that John is in the bar, I can't consistently *disbelieve* the proposition 'He's either in the bar or in the library'; if I have any epistemic attitude to that proposition, it should be one of belief, however inappropriate it is for me to assert it. Similarly for 'You won't eat those and live' when I believe you won't eat them. But the difficulties with the truth-functional conditional cannot be explained away in terms of what is an inappropriate conversational remark. They arise at the level of belief. Believing that John is in the bar does not make it logically impermissible to disbelieve 'if he's not in the bar he's in the library'. Believing you won't eat them, I may without irrationality disbelieve 'if you eat them you will die'. Believing that the Queen is not at home, I may without irrationality reject the claim that if she's home, she will be worried about my whereabouts. As facts about the norms to which people defer, these claims can be tested.[10] But, to reiterate, the main point is not the empirical one. We need to be able to discriminate believable from unbelievable conditionals whose antecedent we think false. The truth-functional account does not allow us to do this.

P. F. Strawson [1986] argues that if Grice is right about indicative conditionals, his thesis should be, and could be, extended to counterfactuals. He gives the examples:

Remark made in the summer of 1964: 'If Goldwater is elected, then

[9] Austin, Ryle, Wittgenstein and others were prone to argue in this way about various important philosophical concepts. The first chapter, 'Prolegomena', of [Grice, 1989] discusses many examples.

[10] I am not talking about cases where you are *certain* that the antecedent is false, which are difficult to assess; but about cases where you think, but are not completely certain, that the antecedent is false.

A good enough test is to take a co-operative subject, who understands that you are merely interested in what she believes, as opposed to what would be a reasonable remark to make; and note which conditionals she assents to.

> the liberals will be dismayed'.
>
> Remark made in the winter of 1964: 'If Goldwater had been elected, then the liberals would have been dismayed'.

and comments that 'the least attractive thing that one could say about the *difference* between these two remarks is that . . . "if . . . then. . . " has a different meaning in one remark from the meaning which it has in the other' (p. 230).

Strawson suggests that the Gricean story can be extended to the second remark. For, if it is made in a context in which it is known that the antecedent is false (or equally, if the form of the remark conventionally suggests that the antecedent is false), then, on the hypothesis that it *is* truth-functional, the hearer is bound to look for some other point to its utterance, and will conclude that the speaker must have just the sort of grounds for it which would have made the first remark reasonable in a context in which the truth-value of the antecedent is not known. All counterfactuals with false antecedents are true; but Grice can explain why some are reasonable things to say, some are not, in terms of principles of good conversation.

Strawson gives no hint that he expects the reader to contrapose: to find the story unbelievable for counterfactuals, and so to weaken its credibility for indicatives too,[11] but this reading is compatible with his conclusion. The truth and meaning of a conditional have now become quite divorced from what matters about it.

Here is a story. English* is identical to English except in one respect which will become clear. It has the word 'dog', and names for various breeds of dog. One breed, however, lacks a name. Speakers habitually call dogs of this breed 'Labradors'. But 'Labrador' really means the same as 'dog'. If you called a poodle a Labrador, this would not be false, but it would be misleading. For you could have said 'poodle'; or, if less specificity is called for, there is the word 'dog', which is shorter and easier to say than 'Labrador'. Hence, speakers tend not to call other breeds 'Labradors'. So the word is quite useful, for when it is used, your audience is likely to cotton on and realise that you aren't speaking of a dog of another breed, but of this nameless one.

The story is incredible. Words mean what people use them to mean, given the distinctions they need to make. Even if 'Labrador' originally meant 'dog', nothing can prevent its coming to perform a more useful role in the language, the name for the nameless breed. Something like this is what Strawson has in mind when he concludes: 'Only in the specially protected environment of a treatise on logic can "$\supset$" keep its meaning pure' (p. 242).[12]

[11] Some post-Dudman readers have already been converted to the view that these forward-looking 'indicatives' behave like 'counterfactuals' (see fn. 5 for references). Then change Strawson's example. *X*: 'If he ate the apple, he was ill'. *Y*: 'He didn't eat it'. *X*: 'Well then, if he had eaten it, he would have been ill'.

[12] My shaggy dog story is a little unfair. Truth conditions for conditionals are problematic, in a way that naming breeds of dog is not. We could be driven to the Gricean manoeuvre as the alternative least at odds with the facts. But it does show that a consistent Gricean story is not necessarily a believable one.

2.6

Frank Jackson has a different defence of the truth-functional account.[13] He claims that there is a special convention governing the assertability of an indicative conditional: it is not enough simply to believe that its truth conditions are satisfied; this belief must be *robust* with respect to the antecedent, that is, it must be that you would not abandon belief in the conditional if you were to discover the antecedent to be true. This ensures that an assertable conditional is fit for modus ponens. This condition is not satisfied if you believe $A \supset B$ solely on the grounds that $\neg A$. If you discovered that A, you would abandon your belief that $A \supset B$ rather than conclude that B.

The details of this defence require the notion of conditional probability to be discussed below (Section 5), and I shall return to Jackson later (Section 9.1). On the face of it, the shift from questions of assertability based on general conversational propriety, to questions of assertability based on a specific convention governing conditionals, leaves the objection to Grice untouched. Jackson speaks of 'the need to facilitate conversational exchanges' [Jackson, 1980, p. 133]. But this doesn't appear to be where the problem with '$\supset$' lies: there is no evidence that one *believes* a conditional whenever one believes the corresponding material implication, and then is prepared to assert it only if some further condition is satisfied.

2.7

Our investigation of the truth-functional conditional leaves us with a conundrum. In Section 2.3 we argued that only truth-functional truth conditions could explain why knowing *just* $A \vee B$ was enough to conclude that if $\neg A$, B; any stronger truth conditions would demand more, and so would not license this inference. In Section 2.4 we argued that the truth-functional account had intolerable consequences, and we have not seen a way to make them tolerable. There is a solution to this conundrum, but it lies ahead.

3 EARLY THEORIES OF COUNTERFACTUALS

3.1

These deserve a mention, for the problems they raise live on. Counterfactuals appeared to be connected not only with dispositional properties but with laws of nature. Laws, it seemed, have counterfactual implications, accidentally true generalizations don't. If we understood counterfactuals, this might illuminate the notion of law. And conversely. Leaving the problem 'What is a law?' for another day, perhaps counterfactuals can be explained as law-governed conditionals. This was

[13] See Jackson [1979; 1980; 1987]; Lewis adopts Jackson's defence [Lewis, 1986, pp. 152–156], having previously [1976, pp. 142–145] supported Grice.

tried by Chisholm [1946], and Nelson Goodman [1947], reprinted as chapter 1 of *Fact, Fiction and Forecast*. Take Goodman's example,

> If the match had been struck, it would have lit.

Its truth, it seemed, requires there to be a law, and facts about the match and its situation (it was dry, there was oxygen, etc.), from which, together with the assumption that it was struck, we can deduce that it lit. A theory of the following shape emerges:

> A counterfactual conditional '$A \rightarrow C$' is true if and only if there is a conjunction of truths T which include a law of nature [and satisfy condition X] such that $A \& T$ entails C.

X is a place-holder for the difficult bit. In fact, the match was not struck, and did not light. Assuming that it was struck involves supposing that some things which are actually true were not true. For instance, the match remained motionless and untouched on the table. True, but this wouldn't have been true if it had been struck, so we need to forget about that fact, in considering what would have happened if it had been struck. But we need to rely upon other things which are actually true, remaining true if the match had been struck, for instance, the fact that it was dry. What distinguishes those facts we may rely upon, from those which we may not, when we make a counterfactual supposition? Using Goodman's name for the problem, which facts are *cotenable* with the assumption that the antecedent is true? (Then the square bracket reads 'and are cotenable with A'.) Goodman defines cotenability thus:

> B is cotenable with A iff it is not the case that if A had been true, B would not have been true. [1955, p. 15]

But now circularity looms: we need cotenability to define counterfactuals and counterfactuals to define cotenability. Consider:

(1) If the match had been struck, it would have lit ($S \rightarrow L$).

(2) If the match had been struck, it would not have been dry ($S \rightarrow \neg D$).

Suppose (1) is true and (2) is false. How does the theory deliver this result? With (1), there is a derivation from the assumption that S, together with a law, and facts such as *it was dry* (D), to the conclusion that L. But with (2), there is a derivation from the assumption that S, together with the same law, and facts such as *it didn't light* ($\neg L$), to the conclusion that it was not dry. The asymmetry must lie in which facts are cotenable with the assumption that it was struck. (1) is true because the match was dry (D), and this is cotenable with the assumption that S; (2) is false because, although the match did not light ($\neg L$), this is not cotenable with the assumption that it was struck. Applying the definition of cotenability, these claims amount to:

> It's not the case that if the match had been struck it would not have been dry [$\neg(S\rightarrow\neg D)$].
>
> It is the case that if the match had been struck it would have lit [$S\rightarrow L$].

Now the circularity is blatant. Why is '$S\rightarrow L$' true but '$S\rightarrow\neg D$' false? Because '$S\rightarrow\neg D$' is false and '$S\rightarrow L$' is true.

Goodman decided that he had reached a dead end. There have been some attempts at escape routes, but no general solution to the problem.[14] Of course, we have an intuitive grasp of what is cotenable with a counterfactual assumption. But then, we have an intuitive grasp of counterfactuals. Goodman was after an explication of that intuitive grasp which did not presuppose it.

3.2

I turn to a different criticism of this approach to counterfactuals (so assume that the problem of cotenability has been solved — or understand that notion intuitively). Is the connection with laws of nature as tight as it requires? Consider:

> If I'd known you were coming, I'd have baked a cake;
>
> If the Labour Party had won, the pound would have fallen;
>
> If you had asked me yesterday, I would have accepted.

Confidence in counterfactuals about our own or others' behaviour, for instance, does not require us to settle the difficult philosophical question whether there are laws of nature from which, together with cotenable facts, the consequent is deducible from the antecedent. But on this account, the counterfactuals stand or fall with the answer to this question. Here is a perfectly ordinary use of a counterfactual: 'They're not at home; for the lights are off, and *if they had been at home, the lights would have been on*'. You might believe this counterfactual even if you are sure that their sitting in the dark is not inconsistent with the laws of nature plus relevant facts. But on this theory, if you are sure that the consequent doesn't follow from laws etc., you should be sure that the counterfactual is false: it deserves zero credibility.

Many of the counterfactuals we accept, about matches, human behaviour, etc., may be roughly on a par with

(3) If you had tossed this (fair) coin ten (or a hundred) times, it would have landed heads at least once.

First, assume indeterminism. Then, on the law-governed account (3) is plainly false: there is no way of deducing consequent from antecedent by law. If you

[14] Bennett [1984] surveys the later literature on the notion of cotenability, pp. 85–8. See also [Bennett, 2003], chapter 20.

know the facts, you know it is certainly false. It deserves zero credibility, and in this respect is indistinguishable from 'If you had tossed the coin ten (or a hundred) times, it would have landed heads every time'. Indeed, it is no more believable than

(4) If you had tossed the coin, it would have turned into a giraffe.

Second, assume determinism. Now, you didn't toss the coin. Nor was it possible for you to do so, given the laws and the past, under determinism. But assume you had tossed it. The antecedent of (3) specifies a type of event that can be realised in many different ways, and the consequent will be sensitive to exactly how you had tossed it. Not all instances of tossing will bring the consequent out true, by the deterministic laws. Again, (3) is plainly false, on this account, and deserves no more credibility than (4).[15] A theory of counterfactuals should explain why, though (3) is not certain, it is plausible and credible while (4) is not.

The coin is merely an illustration of a general difficulty. For many of the things that happen, the disjunction, indeterminism or fine-tuned determinism, is the safest of bets. If we accept it, and the 'law-governed' account of counterfactuals, there is some risk that all contingent counterfactuals whose consequents are at all specific, whose antecedents are not unutterably long and whose consequents are not formulated specifically in terms of chances, turn out false. The explanatory and inferential use we make of such counterfactuals as 'If Mary had asked John to do the shopping, he would have done it', 'If I had climbed over the wall, the dog would have attacked me', 'If Bill had been in London, he would have been in touch' would be vitiated.

4 POSSIBLE WORLDS SEMANTICS

4.1

With Saul Kripke's semantics for modal logic [Kripke, 1963] came the revival of the philosopher's dream, a possible world. It is a promising tool for the elucidation of non-truth-functional sentential connectives. It is certainly useful in the formulation and clarification of modal thought. And it is natural to turn to it for an elucidation of conditionals, which, on the face of it, are *about* possible situations. In the late 1960s David Lewis, Robert Stalnaker and Richmond Thomason developed closely related theories, Stalnaker and Thomason for conditionals in general,

[15] On one interpretation of Davidson's anomalous monism, (see 'Mental Events' in [Davidson, 1980]) counterfactuals with mental content will be like (3) under determinism. 'If you had invited me, I would have accepted'. Events of these mental kinds can be instantiated in various physical ways, and it is under physical descriptions that they instantiate laws. It may be that the laws guarantee a physical realisation of the consequent for many but not for all physical realisations of the antecedent. (There are no strict laws statable in mental terms.) In such a case we would want the counterfactual to come out as probable but not certain, while on Goodman's account it comes out as certainly false.

with only pragmatic differences between indicatives and counterfactuals, Lewis just for counterfactuals. The opening sentence of Lewis's *Counterfactuals* gives the gist:

> *'If kangaroos had no tails, they would topple over'* seems to me to mean something like this: in any possible state of affairs in which kangaroos have no tails, and which resembles our actual state of affairs as much as kangaroos having no tails permits it to, the kangaroos topple over. [Lewis, 1973, p. 1]

Stalnaker says:

> Consider a possible world in which *A* is true, and which otherwise differs minimally from the actual world. *'If* A*, then* B*' is true (false) just in case* B *is true (false) in that possible world.* [Stalnaker, 1968, pp. 33–34]

Between Stalnaker and Lewis, there are differences in formulation, and some substantive differences, but also a difference in aim. Stalnaker's project is less ambitious. He does not expect there to be an informative analysis of '*A*-world which differs minimally from the actual world' which could be specified independently of judgements about what would have been true if *A* were true. Lewis seeks a genuine analysis of counterfactuals in terms which do not presuppose them.[16]

Similarity to the actual world plays the role in these theories which cotenability plays in Goodman's. Goodman's truth conditions, in possible-world jargon, have the form: '$A \rightarrow C$' is true iff in any possible world in which *A* is true and *X* is satisfied, *C* is true. For Lewis and Stalnaker the problem of specifying *X* is the problem of deciding which worlds are closest to actuality.

Similarity is, of course, vague. Comparing cities, or faces, or worlds, there may be no determinate answer to the question: is *A* more similar to *B* than *C* is? But equally, there may be no determinate answer to the question: what would have happened if *A* had been true? Lewis's aim is to analyze one vague notion in terms of another. On the other hand, similarity is not so vague as to be useless. Often, clear judgements can be made about the comparative overall similarity of cities, people, etc., or of how lifelike as opposed to fantastical is a novel or a film.

[16]The analysis of counterfactuals is, for Lewis, part of a larger picture: causation is to be analyzed in terms of counterfactuals, mental states defined as occupants of causal roles, semantic facts obtain in virtue of mental states... His name for the project is 'Humean Supervenience', 'all there is to the world is a vast mosaic of local matters of particular fact, just one little thing and then another' [Lewis, 1986, p. *ix*]. By assuming that there are other possible worlds besides this one, he hopes to be able to reconcile most of what we believe in with an austere view of the fundamental nature of our world.

Lewis's theory of counterfactuals is much more widely accepted than his theory of the nature of possible worlds. I shall say nothing about the latter.

4.2

Lewis's truth conditions for counterfactuals are as follows:

(i) If A is true in no possible world, $A \mathrel{\Box\!\!\rightarrow} C$ is vacuously true.

(ii) $A \mathrel{\Box\!\!\rightarrow} C$ is non-vacuously true if and only if some $A\&C$-world is closer to the actual world than any $A\&\neg C$-world. 'In other words, a counterfactual is non-vacuously true iff it takes less of a departure from actuality to make the consequent true along with the antecedent than it does to make the antecedent true without the consequent'. [Lewis, 1979, p. 164]

If there is a unique closest A-world, $A \mathrel{\Box\!\!\rightarrow} C$ is true iff C is true at the closest A-world. But there may not be: (a) there may be no A-worlds at all, in which case the counterfactual is vacuously true; (b) there may be ties for first place, and C may be true in some but not all of the closest. The literature abounds with examples like:

> If Bizet and Verdi were compatriots, Bizet would be Italian;
>
> If Bizet and Verdi were compatriots, Verdi would be French.

If the closest Bizet-and-Verdi-compatriot worlds contain some in which Bizet was Italian and some in which Verdi was French, then, on Lewis's account, both these counterfactuals are false. (Here he differs from Stalnaker, for whom they have no determinate truth value.) (c) Perhaps there is no closest A-world because for any A-world there is a closer one. Consider a conditional of the form 'If I were taller than I am, C'. Consider a world in which I am an inch taller; then there is a closer world in which I am half an inch taller; and so on, ad infinitum. For Lewis, the conditional is true iff some taller-& -C-world is closer than any taller-& -$\neg C$-world. This has the mildly embarrassing consequence that, given that some differences in height are too small to be detectable, 'If I were taller than I am, no one would know the difference' comes out as incontrovertibly true. Again, suppose you are a little taller than me, say half an inch taller. Then 'If I were taller, I would still be shorter than you' also comes out absolutely certainly and obviously true; whereas, 'Well, maybe, but not necessarily' is a common response to this thought. This example is not very interesting in itself (one like it is mentioned by Lewis in his case against the assumption that there must be a closest A-world); but it serves to illustrate a question to which we shall return: why put all your eggs in the closest baskets?

4.3

Lewis calls the counterfactual a 'variably-strict conditional'. There is the material conditional, $A \supset B$; there is the strict conditional, $\Box(A \supset B)$ — in all possible worlds, $A \supset B$; we could define weaker strict conditionals with reference to some subset of possible worlds, e.g. all those with our laws of nature; but for the

counterfactual, the degree of strictness depends on the antecedent: we depart from the actual world enough to include some A-world; throughout some 'A-permitting sphere' of possible worlds, $A \supset B$ is true. This explains some curious logical properties of counterfactuals. For example, a piece of masonry falls from the cornice of a building, narrowly missing a worker. The foreman says: 'If you had been standing a foot to the left, you would have been killed; but if you had (also) been wearing your hard hat, you would have been all right'; i.e. he says

$$S \ \Box\!\!\rightarrow K; \text{ but } (S\,\&H) \ \Box\!\!\rightarrow \neg K.$$

Strengthening of the antecedent fails for counterfactuals: the nearest S-worlds are K-worlds; but the nearest $S\&H$-worlds are $\neg K$-worlds.

Failures of strengthening are failures of transitivity; for $(S\,\&H) \ \Box\!\!\rightarrow S$ is obviously true; yet we have $S \ \Box\!\!\rightarrow K$ true and $(S\,\&H) \ \Box\!\!\rightarrow K$ false. Other failures of transitivity can be constructed, for instance:

(2) If Brown had been appointed, Jones would have resigned immediately afterwards;

(1) If Jones had died before the appointment was made, Brown would have been appointed; but not:

(3) If Jones had died before the appointment, Jones would have resigned immediately after the appointment.[17]

Departing from reality enough to get Brown appointed has Jones resigning. Departing from it further, to get Jones dead, has Brown appointed. On this reading, (3) does not follow.

It helped to get you to read (2) before (1); if (1) had come first you might have said, after reading (2), 'But not if he was dead!'. Crispin Wright [1983] has argued that the same possible worlds should be in play throughout a single piece of reasoning or discourse (see also [Lowe, 1990]). When they are, transitivity holds. Wright's intuition is mirrored in Lewis's semantics by the validity of

$$A \ \Box\!\!\rightarrow B; (A\&B) \ \Box\!\!\rightarrow C; \text{ so } A \ \Box\!\!\rightarrow C.$$

This restricted transitivity prevents the first premiss from being 'further out' than the second. Wright holds that the '$A\&$' is, as it were, silent, always contextually implied, in the second premiss. Against Wright, the building foreman's remarks above, violating transitivity as they do, constitute a single, pointful piece of discourse; and one can believe both premisses about Brown and Jones. Naturally, if one says something of the form 'If A then B and if B then C', there is presumed to be some point in this utterance, and the most natural one (other than that of producing a philosophical counterexample) is that the hearer is being asked to conclude that if A then C. But Lewis need not deny that.

[17] This example comes from Adams [1965].

It is not as though we should have a bad conscience about all the times we have used or accepted transitive reasoning. First, if the conditionals involved are necessary or a priori, as in maths, logic and sometimes in philosophy, the reasoning does not fail. In other cases the test is whether the second premiss, $B \mathrel{\Box\!\!\rightarrow} C$, survives the addition of A to the antecedent — survives conversion into 'If B [still assuming A], then C'. If it does, the conclusion follows. And it usually does. Wright hypothesises that we always read the second premiss that way. So the same sentence as second premiss will have a different content in different arguments. I don't think there is a deep issue here: we could go Wright's way and save transitivity at the price of increasing ambiguity or context-dependence. But the strengthening case suggests we need not.

There are also failures of contraposition. Stalnaker's example [Stalnaker, 1968, p. 39]:

> If the US had halted the bombing, North Vietnam would not have agreed to negotiate;
>
> but not: If North Vietnam had agreed to negotiate, the US would not have halted the bombing.

And Conditional Proof fails. '$\neg(A\&B)$; A; therefore $\neg B$' is a valid argument form; but '$\neg(A\&B)$; therefore $A \mathrel{\Box\!\!\rightarrow} \neg B$ is invalid. Let A be 'She was hit by a bomb yesterday' and B be 'She was injured yesterday'; it does not follow from the falsity of $A\&B$ that if she had been hit by a bomb, she would not have been injured, i.e. that in the closest possible world in which she was hit by a bomb, she was not injured.

4.4

Laws of nature are not mentioned in Lewis's truth conditions. But he can explain why they loom large in judgements about counterfactuals. Laws of nature are important truths which say much about the character of the world. In general, the difference between two worlds with the same laws will be less than the difference between two worlds with different laws. If, in assessing counterfactuals, we stick as close to the actual world as the specified difference allows, it follows that we tend to consider worlds with the same laws as ours. Thus Lewis explains the connection Goodman took as primitive. And, prima facie, he will have no difficulty with examples like 'If I'd known you were coming I'd have baked a cake'. There is no requirement that the consequent be derivable from the antecedent and premisses including laws.

'Though similarities and differences in laws have some tendency to outweigh differences or similarities in particular facts, I do not think they invariably do so', says Lewis [1973, p. 75]. His reason is as follows. A tree blows over, destroying the roof of a house. Suppose our world is deterministic, at least with respect to the causal chains connected to these events. Consider

If the tree hadn't blown over, the roof would be intact.

Now consider two possible worlds in which the tree didn't blow over. In w_1 the laws are exactly the same as in the actual world. By hypothesis, the relevant laws are deterministic. Then, as the tree didn't blow over, something earlier must have been different, and something earlier still, and so on, back to the beginning of time. This different history has further forward consequences, and w_1, at t, the time in question, is staggeringly different from the actual world.

The history of w_2 is just like the actual world until just before t. Then there is a small, inconspicuous, violation of law — a 'tiny miracle' (relative to the laws of the actual world), and the tree stays upright.

Which of these two worlds is most like the actual world? Surely w_2, the one which does not obey exactly our laws. Lewis writes 'Laws are very important, but great masses of particular fact count for something too.... I therefore proceed on the assumption that the preeminence of laws... is a matter only of degree' [Lewis, 1973, p. 75].

This is the beginning of an apparent difficulty for Lewis's account.[18] His single guiding principle behind counterfactual judgements is overall similarity to the actual world. In opting for w_2 rather than w_1, we keep the past in line, at the price of a 'small miracle'. But the future of the actual world is very different from the future of w_2. Why not purchase future similarity at the price of another small miracle which destroys the roof despite the tree remaining upright? Consider w_3: like w_2, its history is just like the actual world to just before t. A small miracle prevents the tree from falling over. But in w_3 another small miracle — e.g., a lightning bolt — destroys the roof. Its future is very similar to the actual world's — some inhabitants are killed, the family is homeless and impoverished, and further dire consequences ensue. Back in w_2, the family continues its peaceful existence, quite unlike what happens in the actual world. w_2, not w_3, is the way things would have been if the tree hadn't blown over. But w_3, not w_2, is (arguably) the more similar to the actual world — for a reason apparently symmetric with Lewis's reason to prefer w_2 to w_1.

The difficulty is general. It is often the case that if something had happened which didn't, the world would have been very different. Suppose Hitler had died in infancy. Then things would have been quite different in the 1930s and 1940s. But consider the world most similar to the actual world in which Hitler died in infancy. (Here, if you prefer, just focus on the time between antecedent and consequent.) That may be one in which some other child grew up to occupy a virtually identical Hitler-like role. Not that that would have happened, mind you. Imagine two films in which Hitler died in infancy. One of them has a non-Hitler doing all the kinds of thing Hitler did. It strikes you as remarkably like the actual world, almost indistinguishable from the newsreels. The other strikes you as a very plau-

[18] Lewis mentioned the problem [Lewis, 1973, p. 76]. It was pursued in two reviews of *Counterfactuals*, [Bennett, 1974] and [Fine, 1975].

sible account of how the world would have been without Hitler — rather different. Judgements of similarity go one way, judgements about counterfactuals, the other.

Lewis replied to this objection, by specifying which aspects of similarity matter most, on the 'standard resolution of vagueness' for counterfactuals. These are the criteria:

(1) It is of the first importance to avoid big, widespread, diverse violations of law.

(2) It is of the second importance to maximize the spatio-temporal region throughout which perfect match of particular fact prevails.

(3) It is of the third importance to avoid even small, localized, simple violation of law.

(4) It is of little or no importance to secure approximate similarity of particular fact, even in matters which concern us greatly. [Lewis, 1979, pp. 47–48]

To see how these work, return to the tree. We are still operating under the assumption of determinism. There was not perfect symmetry between Lewis's case for w_2 over w_1, and my case for w_3 over w_2. In w_2 the past is *exactly the same* as in the actual world. In w_3, the future is approximately the same as in the actual world; but I did not imagine that the second miraculous disaster would make the world *exactly* as it actually is, with the tree blown over. It would take a massive miracle to secure perfect reconvergence to the actual world, and (1) rules out similar futures at that price. By (2), we prefer w_2 to w_1 as Lewis requires. w_3 has two tiny miracles, w_2 only one, so, by (3), w_2 is to be preferred to w_3, despite the greater approximate similarity of particular fact in w_3, which, by (4), counts for 'little or nothing'. We get the right answer: the most similar world, by these criteria, is the one that would have happened.

Lewis is not a determinist, and in a Postscript to this article [Lewis, 1986, pp. 58–65], he discusses what happens when we drop that assumption. A theory of counterfactuals should not require determinism. Suppose there was some chance that the tree would not blow over. So no small miracle is required to keep the past in line in worlds in which it did not. Lewis puts most effort into arguing that we should discount worlds in which, although the tree doesn't blow over, a 'quasi-miracle' secures *perfect* reconvergence to the actual world. Even if this has a non-zero chance of happening, such peculiar things[19] don't happen in worlds similar to ours, he claims. Let us grant him this. But he is too cavalier about the possibility, which also has a non-zero chance of occurring, of getting the worlds approximately back in line again. He says:

[19] A quasi-miracle is not just a very improbable event. Very improbable events happen in this world and those like it. 'What makes a quasi-miracle is... the remarkable way in which the chance outcomes seem to conspire to produce a pattern [like]... the monkey at the typewriter [producing] a 950-page dissertation on... anti-realism' [Lewis, 1986, p. 60].

> The thing to say about approximate convergence remains the same. Even if approximate convergence is cheap — and even if it is cheaper still when it can be had without even a little miracle — still we can say that it counts for little or nothing, so it is not the case that if Nixon had pressed the button, there would have been approximate convergence to our world, and no holocaust.[20] [Lewis, 1986, p. 59].

Suppose there was a tiny chance at t (but no later than t) of (e.g.) a lightning bolt destroying the roof. It didn't happen: if the tree hadn't blown over, the roof would have been intact. But now w_2 and w_3 minus their miracles, and with the same stretch of identical pasts, are equally suitable by criteria (1)–(3). If approximate similarity counts for nothing, we have a tie, and have been given no guidance on choosing between them. If approximate similarity counts for something, albeit little rather than nothing, then, arguably, the wrong world (w_3) wins. Similarly for the Nixon example.

Consider Kennedy's assassination. Suppose Oswald did it, acting alone, and that if he hadn't, no one else would have. Consider some possible worlds in which Oswald had last-minute fright and did not shoot. Ex hypothesi, no one else even thought of shooting Kennedy. But the crowd contained people carrying guns and not averse to using them who couldn't stand the man: assume, what may well be true, that it was consistent with the laws of nature and the past that someone else act on a sudden impulse to shoot. Again we have two possible worlds in which Oswald didn't do it, not distinguished by Lewis' s criteria (1)–(3), one of which is what would have happened; the other, in which someone else shoots, the more approximately similar to ours. If this counts for nothing, a tie; if it counts a little, similarity takes us in the wrong direction.

Should it be 'little' or 'nothing'? Lewis [1986, p. 48] isn't sure. Many examples suggest that approximate similarity counts for something: 'If I had bet on heads, I would have won'; 'If I had bought these shares last year, I would be rich today'; 'If I had left 5 minutes earlier, I would have avoided the accident'; these do rely on approximate similarity to the actual world after the divergence from perfect match needed to get the antecedent true. There are countless examples like these. To say 'nothing' is to deny the truth of any counterfactual like 'If I had got out of bed one minute earlier, the result of the Swedish election would have been no different'. The example which tempts Lewis to say 'nothing' is due to Pavel Tichy [1976]: when Fred goes out, if the weather is bad, he always wears his hat; if the weather is fine, it's a random 50–50 whether he wears his hat. In fact the weather is bad, and he wears his hat. Consider 'If the weather had been fine, he would have worn his hat'. The fine-weather world in which he does so is more like the actual world than the fine-weather world in which he does not, but the counterfactual is not clearly true.

It is not difficult to spot the difference between Tichy's example and the earlier ones. The weather is not causally independent of whether Fred wears his hat:

[20] This is the example, first used by Fine [1975], in terms of which Lewis conducts this discussion.

fine weather reduces the chance from 100% to 50%. By contrast, my getting out of bed is causally independent of the Swedish election, my buying shares has a negligible effect on their price, etc. But Lewis does not allow himself access to the notion of causation in analyzing counterfactuals, for they are to be used to analyze causation [Lewis, 1973a, pp. 159–72]. His difficulty here generates further doubt about whether the notion of similarity alone, however tailored, will yield the right judgements about what would have been true if *A* had been true. Another doubt about Lewis's criteria was raised by John Pollock (mentioned by Bennett [1984, p. 68]): I leave my coat in a restaurant at noon, and return for it at midnight. A steady stream of potential coat-thieves have passed it by, but it is still there. By Lewis's criterion (2) above, (p. 147), 'If my coat had been stolen this p.m. it would have been stolen very close to midnight' comes out true.

4.5

Lewis's elaborated theory has the effect that we stick to the laws of the actual world at times later than the antecedent-time, *t*, when we evaluate counterfactuals, and actual facts at times later that are unimportant at best. (We exclude worlds with 'quasi-miracles' as well: this doesn't concern me.). Recall that unless the consequent is true in *all* closest antecedent-worlds, the counterfactual is false. Now Lewis's theory is in the same position as Goodman's (see above, Section 3.4). If the consequent is true in almost all close antecedent-worlds, the counterfactual is false, and deserves zero credibility. Again,

(3) If you had tossed the coin ten times, it would have landed heads at least once,

is no more worthy of belief than

(4) If you had tossed the coin ten times, it would have turned into a giraffe.

If indeterminism is rife, almost all counterfactuals about what would happen if you had struck the match, invited me for dinner, etc., turn out false. And if determinism is true but fine-grained, while there is no way that the antecedent *could have* come about, given the laws and the past, the laws won't guarantee the consequent for any old small miracle getting the antecedent true — maybe the vast majority of 'close' ways of instantiating the antecedent will guarantee the consequent but the odd one won't. Then again, all such counterfactuals are false. Contingent counterfactuals, except those with very unspecific consequents ('If you had tossed the coin, it would have landed', perhaps), or consequents about chances, or unutterably long antecedents, will come out false.

Suppose we want the result that someone who knows the relevant facts (be they indeterministic or fine-grained deterministic) should be almost, but not quite, certain that (3). Then we should want something along these lines: a measure of

the credibility of a counterfactual is the proportion of close A-worlds in which C is true. But it is not clear what the truth of a counterfactual like (3) would consist in.

Where there is a tie for closeness and the consequent is true in some but not all closest antecedent worlds, Stalnaker makes the conditional indeterminate — neither determinately true nor determinately false [Stalnaker, 1981, p. 87]. This is more promising from the point of view of the previous paragraph, for it is compatible with a counterfactual like (3) being 'almost true'; whereas, for Lewis, it is 'flatly, determinately false' [Lewis, 1981, p. 331].

4.6

There is a related difficulty for Stalnaker and Lewis, mentioned on p. 252 above (why put all your eggs in the closest basket?). Suppose, on the right account of closeness, a B-world wins among antecedent-worlds, but $\neg B$-worlds are only a hair's breadth behind (as it were). A wins the election. If he hadn't, it would take minimally less departure from actuality for B to win than for C to win. For Lewis and Stalnaker, $\neg A \mathrel{\Box\!\!\rightarrow} B$ is clearly true. It is not even true that C might have won if A hadn't, on either of Lewis's readings of 'might have' [Lewis, 1986, pp. 63–64]. If we find it more acceptable to say that it is only probable that B would have won if A hadn't, we are taking a probability distribution over close $\neg A$-worlds, which is consonant with how, I suggest, we should react to (3). I turn to a theory of uncertain conditional judgements.

5 CONDITIONAL UNCERTAINTY

5.1

You may be sure that B if A, but often you will be less than sure (e.g. that the patient will recover if he has the operation). There are different degrees of uncertainty. You may be nearly sure, fairly sure, think it more likely than not, less likely than not, down to being certain that it's not the case that B if A. The same goes for propositions in general. Our capacity for a spectrum of epistemic attitudes towards a proposition is important in our deliberations about what to do, and what else to think. Often certainty is unachievable, and near-certainty is nearly as good. One application of the theory of probability is to provide a 'logic of partial belief' as Ramsey called it in his 1926 paper 'Truth and Probability' [1931, p. 166]. The theory has its own way of expressing conditional uncertainty — you don't get far in the study of probability without meeting the 'conditional probability of B given A'. In his 1929 paper 'General Propositions and Causality' Ramsey suggested that deliberation about whether if A, B fits the probabilistic model [Ramsey, 1931, pp. 246–47]. In the 1960s and 1970s several philosophers sought illumina-

tion about conditionals from this source.[21]

Any theory of conditionals has consequences for less-than-certain judgements. Something is proposed of the form: 'If A, B' is true iff $A * B$.[22] If a clear-headed person, free from confusions of a logical, linguistic or referential sort, can be nearly sure that $A * B$ yet far from sure that if A, B, or vice versa, then this is strong evidence against the proposal. I am not suggesting that a competent user of 'if' has figured out the correct theory of conditionals. But if a theory states an equivalence between items of belief to which competent users stably, incorrigibly and unhesitatingly take different attitudes (and their practice serves them well), then, on the face of it, the theory is wrong. At best, it has a lot of explaining to do of massive error; and it is hard to see what would convince us, in a case like 'if', that the people are wrong and the theory is right.

We have already seen this pattern of argument at work:

(1) Proposal: 'If A, B' is true iff $A \supset B$, i.e. $\neg A \vee B$. Objection: suppose I'm 90% certain that $\neg A$, hence 90% certain that a sufficient condition for the truth of the right hand side is satisfied, yet 0% certain that if A, B. (Let A be 'The Queen is at home' and B be 'She's worrying about me'. See pp. 135–135 above.) If the proposal were correct, I would be guilty of gross irrationality. But I am not, so the proposal is incorrect.

(2) Proposal: a counterfactual '$A \rightarrow B$' is true iff B is deducible from A + laws of nature + suitable facts. Objection: (a) I may be highly sceptical about whether 'The lights are on' is deducible from 'They are at home' + laws of nature etc., yet close to certain that if they had been at home, the lights would have been on. (b) I am sure that it is not the case that 'The coin landed heads at least once' is deducible from 'The coin was tossed ten times' + laws of nature etc. Yet I am close to certain that if it had been tossed ten times, it would have landed heads at least once. (See above, pp. 140–141.)

(3) Proposal: a counterfactual '$A \boxdot\!\!\rightarrow B$' is true iff B is true at all closest A-worlds. Objection: (a) take the coin example. I am sure that (even if determinism is true) it' s not the case that the consequent is true at all closest antecedent-worlds (but only in the vast majority of them). Yet I am close to certain that if the coin had been tossed ten times, it would have landed heads at least once. (b) I am certain that a world in which I am taller than I actually am and still shorter than you is closer to the actual world than any world in which I am taller than I actually am but not shorter than you. But I am less than certain that if I had been taller, I would still have been shorter than you.

[21] Richard Jeffrey [1964] published an abstract of a paper on this theme. Ernest Adams, [1965; 1966; 1975] has done most to develop this line of thought. See also Brian Ellis [1973] and Robert Stalnaker [1970] for early work on this idea.

[22] Not every theory must take this form. Our present interest is in those which do.

Some degree of uncertainty is the norm for the contingent conditionals we meet every day. It is not a peripheral objection to a theory that it gets uncertain judgements wrong. Can we find a substitution for $A * B$ which is immune to this objection? To find out, we must first turn to the logic of less-than-certain judgements.

5.2

Extend the term 'belief' to include partial belief. If we idealise by expressing a person's degree of belief[23] in a proposition as a number between 1 (for certainty that it's true) and 0 (for certainty that it's false),[24] then it can be shown that satisfying the principles of probability theory is a requirement of consistency[25] upon a person's degrees of belief. Call 'a partition' a set of propositions which are mutually exclusive and jointly exhaustive — not more than one of them can be true, and it must be that one of them is true. The claim that degrees of belief behave like probabilities is the claim that your degrees of belief in the members of a partition should sum to 1. Some consequences of this claim:[26] (1) Your degree of belief in $\neg A$ should be one minus your degree of belief in A (because $\{A, \neg A\}$ is a partition). (2) If you recognise that A and B are incompatible, then your degree of belief in $A \vee B$ should be the sum of your degrees of belief in A and in B (because $\{A, B, \neg A\&\neg B\}$ and $\{A \vee B, \neg A\&\neg B\}$ are both partitions). (3) Your degree of belief in A should be the sum of your degrees of belief in $A\&B$ and $A\&\neg B$ (because $\{A, \neg A\}$ and $\{A\&B, A\&\neg B, \neg A\}$ are both partitions). (4) If you recognise that A entails B, then your degree of belief in A should not be greater than your degree of belief in B. (Write 'b(-)' for your degree of belief in (-). Recognising that A entails B, you have $b(A\&\neg B)= 0$. So $b(A) = b(A\&B)$, by (3). $b(B) = b(A\&B) + b(\neg A\&B) \geq b(A)$.) (5) If you recognise that A and B are logically equivalent, you should have the same degree of belief in each (from (4)).

A partition slices up a space of possibilities. It is convenient to represent applications of the Partition Principle by pictures, as in Figure 1.

The rectangles are of height 1. The internal horizontal lines represent how

[23] I now think 'degree of belief' is not the best name for one's degree of closeness to certainty, given that 'belief' also has an entrenched 'absolute' use such that one definitely does not believe something which one judges to be, say 40% likely. The technical term 'credence' is to be preferred for degree of closeness to certainty. However, I have not altered the text.

[24] For many things that come in degrees, it is a useful idealisation to represent degrees numerically. For then we can use relations between numbers (arithmetic) in our theory of relations between degrees. This — often in physics as well as in philosophy — represents the phenomena as more precise than they really are. Such is the nature of idealisations — but their utility is beyond doubt. The test of the adequacy of an idealisation is that it deliver results of the right order of magnitude.

[25] Consistency is not the only virtue of a set of partial beliefs. But it is all that concerns us here, in investigating the logical relations between degrees of belief.

[26] Here I assume that A and B are bivalent propositions. It will be important later that we need not assume this. We could allow for truth-value gaps, indeterminate or intermediate values, provided we read '$\neg$' as 'It is not true that' rather than 'It is false that'.

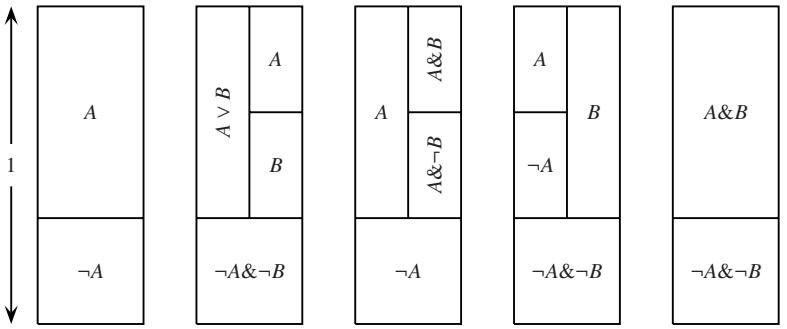


Figure 1.

you divide your belief between the possibilities, in accordance with the Partition Principle.

The other 'fundamental law of probable belief' introduces the idea of a conditional probability:

> (CB) Degree of belief in (p and q) = degree of belief in $p \times$ degree of belief in q given p [Ramsey, 1931, p. 181].

This was not an innovation. Thomas Bayes, in an essay published posthumously in 1763, has as Proposition 3 'The probability that two... events will both happen is... the probability of the first, [multiplied by] the probability of the second *on the supposition that the first happens*' [Bayes, 1940, p. 378], (my italics). The Fourth Principle of Laplace's *Essai philosophique sur les probabilités* (Laplace, 1795, in [Laplace, 1951, p. 14]) is the same.[27] The notion of a conditional probability — the probability that B *on the supposition that* A — plays a big role in many applications of the theory of probability. Now 'on the supposition that A' and 'given A' would appear to be mere stylistic variations on 'if A'. So, it would seem, CB states a logical relation which should hold between your degrees of belief in B if A, $A\&B$, and A. Again writing 'b' for 'your degree of belief in', it may be rewritten

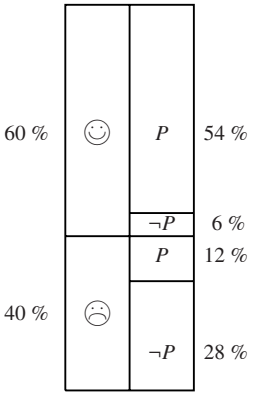
$$b(B \text{ if } A) = b(A\&B)/b(A).$$

[27] Nor was it an innovation to think of probabilities as degrees of epistemic uncertainty. The title of Bernoulli's famous work (1713) is *Ars Conjectandi* — the art of conjecturing. What was new was the argument that degrees of belief have the structure of probability.

Call this **'The Thesis'**.

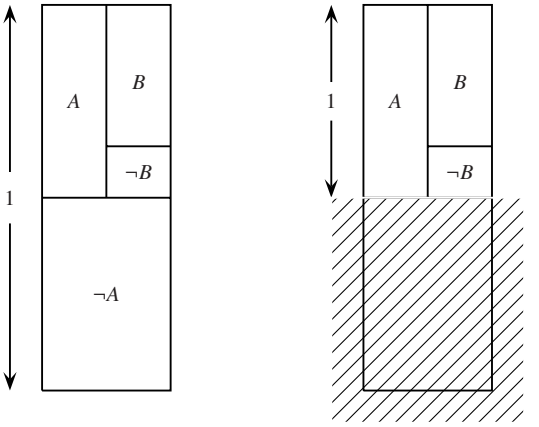
Examples of CB at work: your degree of belief in $A\&B$ is not in general determined by your degrees of belief in A and in B. Suppose you have degrees of belief 1/2 in each of the following: heads on toss 1 (H_1); tails on toss 1 (T_1); not tails on toss 1 ($\neg T_1$); and heads on toss 2 (H_2); yet $b(H_1\&T_1) = 0$, $b(H_1\&\neg T_1) = 1/2$, $b(H_1\&H_2) = 1/4$. The difference lies not in your degrees of belief in the conjuncts, but in the facts that $b(T_1 \text{ if } H_1) = 0$, $b(\neg T_1 \text{ if } H_1) = 1$, $b(H_2 \text{ if } H_1) = 1/2$. Each case is an instance of $b(A\&B) = b(A) \times b(B \text{ if } A)$.

Another illustration: the examiner is to select at random one of five topics for the exam. You are around 90% certain that Jim will pass if one of the three Nice Topics is selected, but only about 30% certain that he will pass if a Nasty Topic (conditionals or probability) is selected. How confident should you be that he will pass? Well, the 60% chance of a Nice Topic divides into: Nice Topic and Pass (90% of 60%); Nice Topic and Not Pass (10% of 60%). The 40% chance of a Nasty Topic divides into Nasty and Pass (30% of 40%); Nasty and Not Pass (70% of 40%). The probability that he will pass is the probability of (Nice Topic and Pass) plus the probability of (Nasty Topic and Pass) which is (90% of 60%) + (30% of 40%) = 54% + 12% = 66%.



These examples used the word 'if', naturally enough, in accordance with the Thesis. When we come to compare the Thesis with rival accounts of 'if', we cannot hijack the word. The standard notation for 'the probability of B given A', understood according to CB, is '$p(B|A)$'. Stressing the interpretation of probability as degree of belief, we may write '$b(B|A)$'. The standard notation (which I will use) is potentially misleading, we shall see, and would be better rendered '$cp(B|A)$', '$cb(B|A)$'. More perspicuous still, perhaps, would be '$p_A(B)$', or '$b_A(B)$'. For your present $cb(B|A)$ is your degree of belief in B, not in your present belief distribution, b, but in a hypothetical belief distribution, b_A, derived from your actual distribution, b, by assuming that A — eliminating the $\neg A$-possibilities — and keeping the

relative probabilities of all the A-possibilities unchanged.



In the pictures, let $b(A) = 0.5, b(A\&B) = 0.4$ and $b(A\&\neg B) = 0.1$. $b_A(B) = b(A\&B)/b(A) = 0.8$. (Note, the Partition Principle applies to b_A as much as to $b : b_A(B) + b_A(\neg B) = 1$; if B and C are incompatible, $b_A(B \vee C) = b_A(B) + b_A(C)$, etc.)

So we have a substantive Thesis about what it is for you to be more or less confident that B if A. You assume A. Under that assumption, you judge it more or less likely that B. And this judgement is equivalent to your judgement of the relative likelihood of $A\&B$ and A. Your degree of belief in an unconditional proposition, that it will rain tomorrow, is proportional to your relative confidence in rain as opposed to no rain: if you think it 9 times more likely that it will rain than that it will not, your degree of belief in rain is 0.9. Your conditional degree of belief that the party will be cancelled (C) if it rains (R), is proportional to your relative confidence in $R\&C$ as opposed to $R\&\neg C$: if you think it 9 times more likely that it will rain and the party will be cancelled, than it is that it will rain and the party won't be cancelled, your degree of belief that it will be cancelled if it rains is 0.9. If you are sure that B if A, e.g. that it has 4 sides if it's square, then $b(A\&B) = b(A)$ and $b(A\&\neg B) = 0$; your degree of belief in $\neg B$ if A is 0. You are nearly sure to the extent that $b(A\&B)$ and $b(A)$ are close, and $b(A\&\neg B)$ is a small fraction of $b(A)$ and of $b(A\&B)$.

Ramsey suggested the Thesis:

> If two people are arguing 'If p will q?' and are both in doubt as to p, they are adding p hypothetically to their stock of knowledge and arguing on that basis about q; ... they are fixing their degrees of belief

in q given p. [Ramsey, 1931, p. 247]

5.3

Further features of conditional degrees of belief need comment.

(1) The ratio $b(A\&B)/b(A)$ is not defined when $b(A)= 0$. It is plausible that the indicative conditional is used only if the antecedent is taken as an epistemic possibility, as not certainly false, by the speaker (thinker) — at least for the sake of argument, at least temporarily, at least to co-operate with her audience. So this is, in the first instance, an exercise in understanding indicative conditionals. Certainty being a vague and shifty notion,[28] there are few things that you cannot take as an epistemic possibility, as all the famous sceptical hypotheses show. Descartes searched for such things, and his findings lend some support to this restriction on indicative conditionals. A conditional may begin 'If I had not existed', or 'If I don't exist tomorrow'. 'If I did not exist yesterday' may get off the ground in the context of a discussion of scepticism; but there is no thought which begins 'If I don't exist now': this is a non-starter.

Adams [1975, Ch. 4] and Brian Skyrms [1981; 1994] suggest (in slightly different ways) that the Thesis can be extended to counterfactuals, along the following lines: confidence in the counterfactual expresses the judgement that it *was* probable that B given A, at a time when A had non-zero probability, even if it no longer does; and even if you do not now have a high degree of belief in B given A.

Probabilities change with time, as live possibilities get eliminated. Think of your favourite thriller: the hero is doomed, escapes with amazing luck, victory seems assured when luck switches to the villain. You bet on 3 heads in a row. Your probability of winning is $\frac{1}{8}$. After one toss, your probability of winning has changed — it is either $\frac{1}{4}$ or 0, depending on the outcome of the first toss. We make judgements about what was probable — was to be expected, was to be expected if we assume such-and-such — as well as about what is probable.

If the Thesis applies to indicative conditionals, an extension to counterfactuals is prima facie desirable. The close links between conditional judgements of the different forms strongly suggest that what makes you confident that (e.g.) you will be ill if you eat the apple, also makes you confident, after you have thrown it away, that you would have been ill if you had eaten it — the conditional probability of illness given eating *was* high. Returning to another example (p. 132 above), the doctor observes certain symptoms. Her degree of belief that the patient has these symptoms is roughly 1; and the assumption that the patient took arsenic has no effect on it: she thinks that if the patient took arsenic, he has these symptoms; and if the patient didn't take arsenic, he has these symptoms. But she thinks that the conditional probability of symptoms given arsenic *was* high, while the conditional probability of symptoms given no arsenic was low; that is, she thinks it was likely

[28] I mean, there is no sharp context-free distinction between certainty and its near neighbours.

that the patient would get these symptoms, given that he took arsenic; and was unlikely that he would get these symptoms, given that he did not take arsenic. She infers from these judgements that the patient probably took arsenic. Returning to yet another example: I can think it was improbable, before the killing that anyone other than Oswald would kill Kennedy. I agree with the judgement which, before the killing, would be expressed by 'If Oswald doesn't do it, no one else will'. But, in the light of what is now known, I am sure that someone else did it if Oswald didn't.

But this is looking ahead. We shall restrict attention to indicatives for the time being.

(2) In mathematical expositions of probability one reads '$p(B \text{ given } A) =_{\text{df}} p(A\&B)/p(A)$ (provided $p(A) \neq 0$), and some philosophers (for example, Lewis [1976, p. 133]) follow suit. A mathematical exposition will start with a complete probability distribution over a partition — an assignment of numbers to the members or 'worlds',[29] which sum to 1. The probability of any (unconditional) proposition is the sum of the probabilities of the worlds in which it is true. The distribution determines $p(A)$ and $p(A\&B)$, in terms of which $p(B|A)$ is defined. This is fine mathematically, but it is at best misleading in epistemic applications of the theory; for it suggests that you need to have determined $b(A)$ and $b(A\&B)$ in order to arrive at $b(B \text{ given } A)$. That would prevent us from working out $b(A\&B)$ from $b(A)$ and $b(B|A)$. In the example of the exam (p. 154) — hardly untypical in structure — we ended up with a partition, constructed from the inputs b(nice), b(pass|nice) and b(pass|not nice). Ramsey's multiplication rule, CB, would collapse into the identity, $b(A\&B) = b(A\&B)$, using this 'definition'. Also, we often have a degree of belief in B given A when we have not determined what we think about A. One important case is when I am deliberating about what to do: A has the form 'I do x', and B is a possible consequence of doing x. It would be absurd to hold that I have to figure out how likely it is that I will do x, before I can arrive at a judgement $b(B|A)$.

The natural order of human thinking is not the same thing as the most elegant order of mathematical exposition.[30] Humans are not endowed with complete belief-distributions over the finest partitions they need to consider. They need to work out some degrees of belief (as the need arises) in terms of others which are more readily accessible. $b(B|A)$ can be accessible en route to $b(A\&B)$, and can

[29] It is convenient to think of the elements of a partition — the finest distinctions among possibilities which are needed for the purpose at hand — as 'worlds'. They are not *ultimate* possibilities — not *complete* ways the world might be, however. We are not capable of thinking of possibilities in complete detail. They are what Kripke calls 'mini-worlds' in the Preface to *Naming and Necessity*, [Kripke, 1972, pp. 16–18].

[30] Russell, in the introduction to the second edition of *Principia Mathematica*, takes the reduction of all truth-functions to the Sheffer stroke to be 'the most definite improvement... during the past fourteen years [since the first edition]' [Russell and Whitehead, 1962, p. *xiii*]. But 'Neither (neither P nor P) nor (neither Q nor Q)' is hardly epistemically more basic than 'P and Q'. Nor could one come by the truth that 2+2 = 4 via the theorem to that effect in Volume 2 of *Principia Mathematica.*

be accessible when $b(A)$ is not (for humans, fortunately, are capable of supposing). The same must be said of another 'definition': A and B are probabilistically independent iff $p(A) = p(A \text{ given } B)$. This does not prevent me from judging that rain in China tomorrow is independent of my finishing this paper next week without having first decided what these probabilities, conditional and unconditional, are and then noticing that they are equal. Whether we see epistemic probability theory as the logic of uncertainty (as I do) or as the mechanics of cognition (as some do) it does not work 'from the bottom up':[31] judgements of independence constrain judgements of unconditional and conditional probability and judgements of conditional probability fix ratios of unconditional probabilities.[32] It is perfectly possible to operate with the constraint that (unless or until something should change my mind) $b(A\&B) = 0.9b(A)$ in advance of settling $b(A)$. The Thesis says that doing so is tantamount to being 90% certain that if A, B. The thought process behind this judgement: assume A; under that assumption, I'm 90% certain that B.

A ratio can be determinate whose numerator and denominator are not. I look at my speedometer, which tells me, in its inscrutable way, that I am now doing 30 miles per hour. 'Now' refers not to a dimensionless instant, but to a short but indeterminate stretch of time. Velocity is distance divided by time. We do not have to resolve the indeterminacy of the time — for however we do so, the ratio is fixed (within limits) and meaningful. Conditional degree of belief is an interesting concept to the extent that the ratios are stable fixtures of a belief system, which can be settled independently of $b(A)$ and $b(A\&B)$.[33]

(3) *A* feature of the Thesis to which some people object[34] is that belief that $A\&B$ is sufficient for belief that B if A. The Thesis shares this feature with the truth-functional conditional, and with Stalnaker's possible-worlds analysis of indicative conditionals: if you believe $A\&B$, you believe a sufficient condition for the truth of $A \supset B$; and you believe that the A-world which differs minimally from the actual world — viz., the actual world itself — is a B-world. Still, it is commonly complained, conditionals with parts which are mutually irrelevant, like 'If Napoleon is dead, Oxford is in England' are not acceptable, or even, false.

Our interest in conditionals centres on the case where we're not sure whether A, not sure whether B, but the supposition that A has some bearing on whether B. I'm not sure whether Jim will pass, but pretty sure that if a Nice Topic comes up he will pass. But let us consider the consequences of the Thesis for the less interesting cases. First, suppose you are already sure that A. Then supposing that

[31] Think of a 'world' as a state-description. Suppose there are six logically independent propositions to be considered in a given problem. The 64 state-descriptions (or lines of a truth table) of 6 conjuncts form a partition. 'From the bottom up' means that we start by assigning probabilities to these.

[32] Pearl [1988] does much to make the theory computationally tractable as the mechanics of cognition, by giving beliefs about independencies a fundamental role.

[33] D. H. Mellor [1993] dismisses the ratio as an account of conditional belief on the grounds that $b(B \text{ if } A)$ can exist when $b(A)$ does not. This is an over-reaction. No defender of the Thesis thinks $b(A)$ must be fixed before $b(B \text{ if } A)$.

[34] For instance Mellor [1993], Pendlebury [1989], and Read [1995].

A changes nothing: your belief distribution b already rules out $\neg A$, and so is the same as b_A. Your degree of belief in any proposition, B, on the assumption that A, is simply your degree of belief in B. So, if you are already sure that B as well as A, you are sure that B if A. Most instances of this kind will be of no interest. But it is too much to ask that all acceptable conditionals be interesting. It is enough that you do not doubt that B is true, on the assumption that A is.

Suppose you think that B is true. It does not follow that for any supposition, A, you will believe that B if A. For A might be the sort of supposition which would undermine your belief that B. But if you consider A to be irrelevant to B, the supposition that A leaves your belief that B undisturbed. For instance, I believe that the match will be cancelled; for all the players have flu. I believe that the match will be cancelled whether or not it rains. I think it will be cancelled if it rains, and I think it will be cancelled if it doesn't rain. (Saying 'The match will be cancelled if it rains' is likely to be misleading in this situation. To reject Grice's defence of the truth-functional conditional is not to reject wholesale the Gricean thought that you can mislead your audience by expressing a belief, when there is something more appropriate you could have said.) On the other hand, although I believe that the match will be cancelled, I don't believe that if the players make a very speedy recovery the match will be cancelled. For that supposition does unsettle my belief.

Mellor [1993] defends a close cousin of the Thesis, but jettisons this feature: he does not, he says, accept 'If France is big, Egypt is hot' (although he is certain of both conjuncts). 'I am not at all disposed to infer Egypt's heat from France's size' (pp. 247–8). Being disposed to infer B from A is one way he characterises accepting a conditional, which he elaborates: 'In other words, fully to accept ... 'If P, Q' is to be disposed fully to believe Q if I fully believe P'. He also endorses Ramsey's explanation of conditional degrees of belief in terms of conditional bets: 'My choice of odds for... a conditional bet [on Q, conditional upon P] I take to measure the degree of belief I now believe I am disposed to have in Q if I fully believe P' (p. 234, fn. 5). But if betting tests work at all, they will show Mellor believing 'If France is big, Egypt is hot'. A conditional bet is a bet on the consequent, which is called off if the antecedent is false. He is sure, in this case, that the antecedent is true, and hence that the bet won't be called off. It is, in his eyes, equivalent to a bet on the consequent. His choice of odds will reflect his belief that the consequent is true, and his belief that it is true if the antecedent is. So Mellor does not have a consistent position on this issue. Elsewhere in his argument, Mellor says he accepts, speaking of a visibly blue bird, 'If that's a canary, it's not yellow' (p. 245). So, I think, he should. But he is no more (or less) disposed to infer 'It's not yellow' from 'It's a canary' than he is to infer 'Egypt is hot' from 'France is big'. On a liberal interpretation, 'to accept a conditional is to be disposed to infer consequent from antecedent' does apply here. Your other beliefs must, in general, be appealed to in these inferences. But then, if Q is one of your other beliefs, Q follows from the assumption that P together with your other beliefs.

Relevance is a context-dependent matter. Any two logically independent propositions are mutually relevant in some contexts, and mutually irrelevant in others.

Thomson [1990] has the example of someone coming home saying 'If there's a book on my coffee table, two Great Danes arrived at Paddington Station this morning' — and tells a story in which the remark is apposite. The relevance of the symptoms to the question whether I have the disease ceases after more direct tests have been carried out. The relevance of antecedent to consequent in a contingent conditional ceases when the truth value of the consequent is established by perception. A defender of the Thesis (or the truth-functional conditional, or Stalnaker's conditional) will claim that questions of relevance belong to the pragmatics of communication. This is, I think, what Lewis calls a 'spoils to the victor' issue: if the best overall theory allows that there are boring but acceptable conditionals with mutually irrelevant parts, so be it. If not, not.

(4) It would be wrong to read 'It is probable that B given A' as 'If A, then (it is probable that B)'. This would be like the so-called 'modal fallacy' — of reading 'If he's sitting down, then necessarily, he's not standing up' with 'necessarily' qualifying the consequent rather than the whole thought. The modal fallacy has the consequence that all truths are necessary truths: if A, then necessarily A. And in this context it has the consequence that all probabilities are 1 or 0. For the probability of A given A is 1, and the probability of A given $\neg A$ is 0. If we read this: if A, then $p(A) = 1$; if $\neg A$, then $p(A) = 0$; then, granted $A \vee \neg A$, we could validly derive that $p(A) = 1$ or $p(A) = 0$. 'I'm sure that A if A' does not have the consequence that if A (is true), then I'm sure that A (is true).

It is less of a howler to think of your degree of belief in B given A as the degree of belief you would have in B if you were certain that A. This is typically correct, but not invariably so. For there are all sort of ways you might learn that A. You think the match will light if struck. You learn that it is being struck. Typically, you then think it will light. But not if you learn that it is being struck at the bottom of a swimming pool.

Admittedly, in this last example, you did not expect it to be struck in this way, or else you wouldn't have thought it would light if struck. So perhaps thinking that B if A is being presently disposed to believe B if you learn that A. Mellor [1993] suggests this.[35] But this won't quite do either — there is one particular kind of counterexample:

> If Reagan was in the pay of the KGB, we'll never find out.

Suppose Reagan was in the pay of the KGB; then, I judge, it's likely that we'll never find out. But if I were to learn that he was in the pay of the KGB, I would not think it likely that we'll never find out! Nor, *pace* Mellor, am I presently disposed to believe the consequent on learning the antecedent. What the example shows is that supposing that something is true is not always equivalent to supposing you know it's true, or pretending you're certain that it's true.[36]

[35] Mellor does not call this a judgement of conditional probability, reserving that name for something he distinguishes from it and rejects. See fn. 33 above.

[36] W. V. O. Quine [1966, pp. 22–3], makes this distinction in his solution to the surprise examination

'The probability that (your degree of belief in) B on the supposition that A' is intelligible as it stands: these further attempts to gloss it are unsuccessful and unnecessary.

Another idiom comes in handy. Take the special case where you have a partition, fine enough for the problem at hand, of equally likely alternatives, or 'worlds'. The probability of an unconditional proposition (A, $\neg A$, $A\&B$, etc.) is the proportion of worlds in which it is true. The probability of B given A is the proportion of A-worlds which are B-worlds. (The proportion of $A\&B$-worlds is the proportion of A-worlds, multiplied by the proportion of A-worlds which are B-worlds.) If we drop the simplifying assumption that each alternative is equally likely, we have to replace 'proportion' by 'weighted proportion', where the weights are the probabilities of the worlds. Just focusing on a single probability distribution, we can stick to the simpler idiom by artificially subdividing the weightier worlds into slimmer ones, indistinguishable for the purposes at hand, until each proposition is true at some number of equally likely worlds.[37] We mirror the structure of conditional and unconditional probabilities by the phrases 'proportion of A-worlds which are B-worlds' and 'proportion of worlds which are A-worlds'.

6 THE BOMBSHELL[38]

6.1

At the beginning of Section 5 we considered a form of objection to various proposed truth conditions for conditionals: that a clear-headed person could have less confidence in the conditional than in the satisfaction of the proposed truth condition, or vice versa. At the end of Section 5.1 we asked: can we find a truth condition which is immune from this objection? We turned to a theory of uncertainty, and arrived at the Thesis: a logical relation which governs your degrees of belief in B if A, $A\&B$, and A. Can we find truth conditions for conditionals[39] which fit the Thesis? Take any two logically independent propositions, e.g. 'Ann is in Paris' and 'Bill is in Paris'; call them A and B; suppose you have a conditional degree of belief in B given A. Which truth conditions (if any) are such that your degree of belief in their obtaining must match your conditional degree of belief in B given A? Can we find a proposition X such that, in any consistent belief distribution over the relevant propositions in which $b(A) \neq 0$,

---

paradox. Examples like this are due to Thomason (see [van Fraassen, 1980, p. 503]). This example is Lewis's [Lewis, 1986, p. 155].

[37] To the diagram on p. 155, superimpose as many equally-spaced horizontal lines as you need to get each member of the partition true in some number of elements of the resulting partition of equally-likely 'worlds'.

[38] So described by Bas Van Fraassen [1976, p. 273]; and by Stalnaker in a letter to Van Fraassen published therein, p. 302.

[39] We are concerned here with indicative conditionals (see pp. 156–157 above). The possibility of extending the Thesis to counterfactuals is discussed in Section 10.

$b(X) = b(B \text{ given } A)$ [henceforth, the Equation]

so that we may say, consistently with the Thesis, 'If A, B' is true if and only if X?

There can be different epistemic attitudes to the same proposition. We seek an X such that $b(X)$ and $b(B|A) = b(A\&B)/b(A)$ cannot coherently come apart: an interpretation such that for all consistent distributions of belief over the relevant domain,[40] $b(X) = b(A\&B)/b(A)$. In a given belief distribution, there may of course be a proposition C (or several such propositions) which you believe to the same degree as you believe B given A. But if someone else, or you in a different information state, may consistently think C more, or less, likely than B given A, C is not the X we seek.

The bombshell is that no proposition at all satisfies the Equation. If we stick by the Thesis, we must not think of conditionals as propositions, as truth bearers. If belief that if A, B fits the Thesis, it is nonsense even to say things of the form " 'If A, B' is true if and only if, if A, B". Your degree of belief that B is true, on the supposition that A is true, cannot be consistently and systematically equated to your degree of belief that something is true, simpliciter.

Some pre-bombshell writings foretold this result. Gilbert Ryle [1950] recommended thinking of conditionals as 'inference tickets' rather than statements. John Mackie [1973, p. 93] construed saying 'If A, B' as asserting that B within the scope of the supposition that A, and said this view 'abandons the claim that conditionals are in a strict sense statements,... that they are in general simply true or simply false'. Adams [1965, pp. 169–170; 1966, pp. 265–266] expressed doubts about the application of truth to conditionals, and developed a logic for conditionals construed in accordance with the Thesis.

But there was no strong reason for holding that there must be an opposition between (e.g.) asserting that B under the supposition that A, and saying something true. Indeed, it was (and is) hard to see how there could be an opposition: must there not be a distinction between when it is right, and when it is wrong, to assert that B under the supposition that A, which will yield a notion of truth and falsity?

Stalnaker [1968] fully endorsed Ramsey's account of conditional belief: add the antecedent hypothetically to your stock of beliefs, and consider whether you believe the consequent under that hypothesis. He sought appropriate truth conditions to match:

> Now we have found an answer to the question, 'How do we decide whether or not we believe a conditional statement?' the problem is to make the transition from belief conditions to truth conditions; that is, to find a set of truth conditions for statements of conditional form which explains why we use the method we do use to evaluate them. The concept of a possible world is just what we need to make this transition, since a possible world is the ontological analogue of a stock

[40] I shall take the qualification 'in which $b(A) \neq 0$' for granted except when it is specially important.

> of hypothetical beliefs. The following set of truth conditions, using this notion, is a first approximation to the account I shall propose:
>
> Consider a possible world in which *A* is true, and which otherwise differs minimally from the actual world. 'If *A*, then *B*' is true (false) just in case *B* is true (false) in that possible world. [Stalnaker, 1968, pp. 33–34]

This was the first appearance in print of the 'nearest possible world' approach to conditionals, and it was designed to provide the 'ontological analogue' of the Thesis. Stalnaker left probabilistic considerations aside in this introductory paper, announcing in a footnote (n. 17, p. 43) that these would be elaborated subsequently [Stalnaker, 1970]. But they were in the background. The same footnote refers to Adams [1966], who had shown, for instance, that there are plausible counterexamples to transitivity, strengthening and contraposition, and that this was to be expected for conditionals which satisfied the Thesis. (The logic Adams developed on the basis of the Thesis is discussed below in Section 7.2.) Stalnaker gave his own counterexamples to these inference patterns [Stalnaker, 1968, pp. 38–39], and formulated a logic which is identical to Adams' over their common domain. (Adams' logic is restricted to sentences in which 'if', if it occurs at all, occurs as the main connective. Stalnaker's is not so restricted: once we have truth conditions, we have something which embeds naturally in longer sentences.)

Coincidence in logic does not guarantee coincidence in interpretation. But Stalnaker's attitude to interpretation was minimalist. He did not expect a reductive analysis of 'if' in terms of a substantive notion, 'minimally different *A*-world'. Rather, we should think of the 'minimally different *A*-world' as the world that will be actual if *A* is. He later described his conditional propositions as 'a projection of epistemic strategy onto the world' [Stalnaker, 1984, p. 119].

Allan Gibbard [1981, p. 211] describes it as 'little more than a coincidence' that Adams' and Stalnaker's logics agree. In one way it is no coincidence: each is motivated by the same notion of conditional belief, which Stalnaker's truth conditions were intended to fit. Nor is it so surprising that the rich framework of possible-worlds semantics, in the hands of someone as expert as Stalnaker in its manipulation, should yield the right structure. With agreement in logic, and no leverage on the semantics independently of the notion of conditional belief, Stalnaker's claim, to have identified the proposition whose belief conditions fit the Thesis [Stalnaker, 1970, p. 107, p. 120], had an irrefutable air. So it came as a bombshell when Lewis, at the 1972 meeting of the Canadian Philosophical Association, refuted it, proving that there is no proposition at all such that your degree of belief in its truth systematically matches your degree of belief in *B* given *A* [Lewis, 1976; 1986a]. A conditional degree of belief is not equivalent to a degree of belief that [something or other] is true.

6.2

There are four bombshells coming up. Here is the base result. There is no proposition X such that $p(X) = p(B|A)$ in all probability distributions in which these are defined. (*A* probability distribution is an assignment of non-negative numbers to the members of a partition which sum to 1.)

Suppose there is such an X. We first show something of the logical relationships between X and A: X is (a) compatible with A, and (b) compatible with $\neg A$, but (c) not entailed by $\neg A$, i.e., X may or may not be true if $\neg A$ is true.

Proofs, in reverse order: (c) There are probability distributions in which $p(\neg A)$ is high and $p(B|A)$ low (e.g., let $p(\neg A) = 0.9$; $p(A\&B) = 0.01$; $p(A\&\neg B) = 0.09$. $p(B|A) = 0.1$.) So there are probability distributions in which $p(\neg A)$ is high and $p(X)$ [$= p(B|A)$] low. So $\neg A$ cannot entail X: if it did, X would be true throughout the $\neg A$-worlds, and could not be less probable than $p(\neg A)$. So, in some $\neg A$-worlds, X is not true.

Similarly: (b) there are probability distributions in which $p(B|A)$ is high and $p(A)$ is low; hence in which $p(X)$ is high and $p(A)$ is low. So X cannot entail $A : X$ must be true in some $\neg A$-worlds. And a parallel argument will show that (a) X cannot entail $\neg A : X$ must be true in some A-worlds. There is nothing surprising in these facts. Now to the main part of the proof:

(i) $p(B|A)$ *depends only on how probabilities are distributed in the A-worlds* (the part of the partition in which A is true). Fix $p(A)$ and $p(A\&B)$, and $p(B|A)$ is fixed.

(ii) Any proposition X which satisfies the Equation must be true in some but not all $\neg A$-worlds, and true in some A-worlds, as was shown. So $p(X)$ depends *not only* on how probabilities are distributed in the A-worlds, but also on how they are distributed in the $\neg A$-worlds.

(iii) There are distinct probability distributions which agree in all assignments in the A-worlds, but disagree in assignments in the $\neg A$ worlds. They will agree on $p(A\&B)$ and $p(A)$, and hence on $p(B|A)$. And they will agree on $p(A\&X)$. But they will disagree on $p(\neg A\&X)$. As $p(X) = p(A\&X) + p(\neg A\&X)$, they will disagree on $p(X)$. So there are distributions in which $p(B|A) \neq p(X)$. End of proof.[41]

We can illustrate the proof by taking Stalnaker's conditional ('$A > B$') for X. We get a partition of logical possibilities of the following shape.

[41] A sketch of a proof along these lines is given by I. Carlstrom and C. Hill [1978], in their review of Adams [1975].

	A	B	$A > B$	p_1	p_2
1.	T	T	T	0.4	0.4
2.	T	F	F	0.1	0.1
3.	F		T	0.4	0.1
4.	F		F	0.1	0.4

(Note: (a) If A is true, the minimally different A-world is the actual world; this explains the top two lines; (*b*) if A is false, then, whatever the truth value of B, it may or may not be the case that the minimally different A-world is a B-world. This explains lines 3 and 4. (c) Appearances to the contrary, we need not assume that the proposition X (here $A > B$) is necessarily either true or false. What is at issue is whether there is a proposition the probability of whose *truth* is, in all distributions, $p(B|A)$. We can leave open whether, whenever it is not true, it is false. Hence we could construe, here and below 'F' as 'not true', and '¬' as 'It is not true that'.)

On the right we have two probability functions over the partition. They agree in the A-worlds. In each $p(B|A) = 0.4/0.5 = 0.8$. In the first, $p_1(A > B) = 0.4 + 0.4 = 0.8 = p_1(B|A)$. In the second, $p_2(A > B) = 0.4 + 0.1 = 0.5 \neq p_2(B|A)$.

Does the base result refute Stalnaker's claim? An argument that it does would be this. We have above four clearly specified possibilities, one and only one of which will obtain, like a 4-horse race. Allow background information to vary. Any probability distribution over four such possibilities might represent a not-irrational belief distribution in some state of background information. (A crude illustration: as I ponder the four possibilities above, an Oracle tells me 'I'll give you a hint: either 2 or 3 is the true one'. Accepting the hint, I divide my belief equally between 2 and 3. If I am certain that the Oracle spoke truly, my $b(B|A)$ and $b(A>B)$ are respectively 0 and 0.5. If I am nearly certain that the Oracle is right, they are close to these numbers.) This argument will be blocked if it can be shown, in a non-question-begging way, that for some probability distributions (the conflicting ones) there is no state of information in which they would represent a reasonable belief-distribution. I do not know any such argument.

But there is an interpretation of Stalnaker which is immune to the base result. He could be interpreted as *stipulating* that, as well as satisfying the partition principle, belief distributions involving conditional propositions and their parts must satisfy the Thesis: p_2 is to be ruled out from the class of consistent belief-distributions. This fits with Stalnaker's image of a conditional proposition as a 'projection of epistemic strategy onto the world'. The fallacy in the argument above was, I suppose, to take 'conditional proposition' too realistically: there are no facts about 'nearest A-worlds' independently of our epistemic strategies.

The stipulation is consistent for a single conditional in a single belief distribution. But stipulations have consequences. This one has untenable ones when we consider the original conditional in different belief distributions (as Lewis showed); it also has untenable consequences for other conditionals in the same belief distribution (as Stalnaker himself showed).

6.3

Here is a simplified and relatively informal version of Lewis's proof (1976). The proof involves an initial belief distribution, which is ex hypothesi reasonable, in which $b(A\&B)$ and $b(A\&\neg B)$ are both non-zero. We find out how it must change (given the Equation) if the believer were to learn certain things, and deduce what it must have been like in the first place for such changes to be rationally permissible. For this last part we need a principle about belief revision. I shall appeal to the following principle, which is weaker than Lewis's, a consequence of his, but all that he needs:

> (RP) If a person rationally has a non-zero degree of belief in $C\&D$; and then learns C for certain, and nothing else of relevance, it is always rational for him to continue to have a non-zero degree of belief in D.

Contraposing the principle: if learning C for certain renders impermissible any degree of belief in D other than 0, then $b(C\&D)$ cannot rationally be other than 0 before you learnt C.

First, suppose the person were to learn for certain that B, and nothing else of relevance. In the new distribution b' that would result, $b'(B) = 1$. As his original $b(A\&B) \neq 0$, $b'(A) \neq 0$ (by RP). So $b'(B|A) = b'(A\&B)/b'(A) = 1$. So $b'(X) = 1$, and $b'(\neg X) = 0$. So, by RP, in the former distribution, it is not rationally permissible to have a non-zero degree of belief in $B\&\neg X$. (If it were, learning B would not force $\neg X$ to zero.)

Second, suppose the person were to learn for certain that $\neg B$, and nothing else of relevance. In the new distribution b'' which would result, $b''(B) = 0$. As $b(A\&\neg B)$ was non-zero, $b''(A) \neq 0$. So $b''(B|A) = b''(A\&B)/b''(A) = 0$. So $b''(X) = 0$. By the same reasoning as above, this means that in his original distribution $b, b(\neg B\&X)$ must have been 0. For if it were not, learning just $\neg B$ would not force him to assign 0 to X.

We have proved that in any reasonable belief distribution in which $b(A\&B)$ and $b(A\&\neg B)$ are non-zero, $b(B\&\neg X) = 0$ and $b(\neg B\&X) = 0$. It is an elementary consequence of these two facts that $b(B) = b(X)$. [$b(B) = b(B\&X)$; and $b(X) = b(B\&X)$.] So, in any such distribution, $b(B) = b(X) = b(B|A)$. But this is absurd! Take any three-way partition e.g. $[C\&D, C\&\neg D, \neg C]$ and a distribution of belief which is positive (e.g. $\frac{1}{3}$) for each member $b(C\&D)/b(C) = \frac{1}{2}, b(C\&D) = \frac{1}{3}, b((C\&D)|C) \neq b(C\&D)$. Yet this example satisfies Lewis's initial conditions; $b(C\&(C\&D))$ and $b(C\&\neg(C\&D))$ are both positive. The absurdity may be stated roughly thus: take any two propositions such that the first does not entail the second; learning that one is true has no relevance to how much you should believe the other.[42]

[42] Anthony Appiah [1986] objected to Lewis's proof, claiming that it is never reasonable to have a degree of belief of strictly 1 in a contingent proposition. Lewis [1986a] gave a new proof which does not require that assumption.

6.4

Van Fraassen [van Fraassen, 1976] objected that it is unreasonable to assume that the sentence 'If A, B' be interpreted as the same proposition in different belief distributions. He labelled this assumption Lewis's 'metaphysical realism' (p. 252). The label is hardly fair in *this* context: the assumption amounts only to the claim that we can take different epistemic attitudes to a proposition without changing the subject (see Lewis [1976, p. 138]). Replying by letter to Van Fraassen in 1974 [1976, pp. 303–304], Stalnaker produced his own version of the bombshell. It is described by Gibbard [1981, pp. 219–220]. We do not need to invoke different belief distributions to derive an absurdity, at least for Stalnaker's proposed truth conditions. He shows that if, in some distribution, $b(A > B) = b(B|A)$, then there exist, in the same distribution, two further propositions, C and D, such that, demonstrably, $b(C > D) \neq b(D|C)$.

Here again, on the left, is a partition for Stalnaker's conditional

	A	B	$A > B$	b	C	D	E	$C>D$
1.	T	T	T	0.25	T	F	F	F
2.	T	F	F	0.25	T	T	F	T
3.	F		T	0.25	F	F	T	F
4.	F		F	0.25	T	F	F	F

Consider a belief distribution which assigns 0.25 to each line, and so satisfies the Equation: $b(A > B) = b(B|A) = 0.5$. (Any numbers other than zeros which satisfy the Equation will do. I just pick the easiest.) Let E be $\neg A\&(A > B)$: E is the proposition true just at line 3. Let C be $\neg E$. C is the proposition true at lines 1, 2 and 4. Let D be $A\&\neg B$. D is true just at line 2. Now $b(D|C) = \frac{1}{3}$ [line 2 (0.25) divided by the sum of lines 1, 2 and 4 (0.75)]. What about C>D? It is true at line 2 (because its antecedent and consequent are); false at lines 1 and 4 (because its antecedent is true and its consequent false). What about line 3? Stalnaker shows that $C > D$ and E are incompatible. So $C > D$ is false at line 3. So it is true just at line 2. So $b(C > D) = 0.25 \neq b(D|C)$. Why are $C > D$ and E incompatible? C has the form $(A \vee (\neg A\&G))$ [I abstract from the structure of G]. So $C > D$ has the form $(A \vee (\neg A\&G) > (A\&\neg B))$. This implies $A > \neg B$,[43] while E implies $A > B$. For consistent $A, A > B$ and $A > \neg B$ are incompatible on Stalnaker's logic and semantics.

Admittedly, 'C>D' is a somewhat contrived proposition. But its existence is

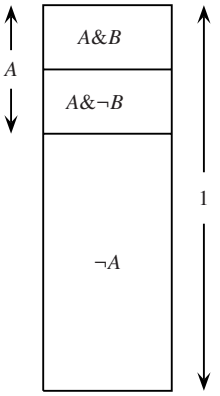
[43] A proof that $C > D$ entails $A > \neg B$, in terms of Stalnaker's semantics: $C > D$ says that the nearest $(A \vee (\neg A\&G))$-world is an $(A\&\neg B)$-world. Suppose that's true. Now the nearest $(A \vee (\neg A\&G)$-world is either an A-world or a $\neg A$-world. Suppose it's a $\neg A$-world. Then the conditional says it's an $(A\&\neg B)$ world. Contradiction. So the nearest $(A \vee (\neg A\&G)$-world must be an A-world. Go to the nearest A-world. It will be the nearest $(A \vee (\neg A\&G)$ world. So it's an $A\&\neg B$-world. So it's a $\neg B$ world. So $A > \neg B$. (The proof, of course, can be done formally in Stalnaker's logic. The only not entirely trivial steps involve (1) the incompatibility of $A > B$ and $A > \neg B$, for consistent A; and (2) that $((A > C) \vee (B > C))$ follows from $((A \vee B) > C)$.)

forced upon us by the assumption that there are truth conditions compatible with the Thesis. (It's not easy to get your mind round the Gödel sentence, either.)[44]

6.5

The following argument yields, I believe, a diagnosis of the trouble. Let us examine the relationship between $b(B \text{ given } A)$ and $b(A \supset B)$. There are two special cases in which they must be equal: (1) you are certain that $A\&\neg B$ is false (but not certain that A is false); then $b(B \text{ given } A)$ and $b(A \supset B)$ are both 1; (2) you are certain that A; then $b(B \text{ given } A) = b(A \supset B) = b(B)$. These cases apart, in all belief distributions $b(B \text{ given } A) < b(A \supset B)$.

The easiest way to see this is to compare how much $b(B|A)$ and $b(A \supset B)$ differ from certainty. Here is a partition. Adjusting the positions of the inner horizontal lines will represent different belief distributions over it.



The amount by which $A \supset B$ differs from certainty is simply the proportion of the whole assigned to $A\&\neg B$: writing 'u' for 'the uncertainty of', $u(A \supset B) = (1 - b(A \supset B)) = b(A\&\neg B)$. The amount by which $b(B|A)$ differs from certainty is the proportion of A which is assigned to $A\&\neg B$: $u(B|A) = (1 - b(B|A)) = b(A\&\neg B/b(A))$. Now, $b(A\&\neg B)$ is a greater proportion of $b(A)$ than it is of the whole — except when $b(A\&\neg B) = 0$, or $b(A) = 1$. Hence $b(B|A)$ is more uncertain than $b(A \supset B)$, except in these two special cases, where they are equal. If $\neg A$ is large, $A\&\neg B$ must be small; but $A\&B$ may be smaller still, in which case $b(B|A)$ is

[44] A more general proof that, within a single belief distribution, not all conditional probabilities can be probabilities of the truth of a proposition, is given by Alan Hájek [1989; 1994].

low but $b(A \supset B)$ is high. An example of the difference: how likely is it that if this (fair) die lands an even number, it will land 6? b(six given even) = $\frac{1}{3}$, b(even $\supset$ six) = b(not even, or six) = $\frac{2}{3}$: if it lands 1, 3, 5 or 6, the truth-functional conditional is true.

When we try to equate $b(B|A)$ with a degree of belief in a proposition, $b(X)$, we find we have incompatible requirements upon it:

(1) Clearly, $A \supset B$ doesn't entail X. (If it did, one could not coherently have a higher degree of belief in $A \supset B$ than in X; but in general, $b(A \supset B) > b(B|A) = b(X)$.) So there are possible situations in which $A \supset B$ is true and X is not true. Hence, someone with just enough information to be certain that $A \supset B$ does not have enough information to be certain that X: ruling out just those situations in which $A \supset B$ is false, i.e. ruling out just $A\&\neg B$, leaves open the possibility that $A \supset B$ is true and X is not.

(2) But, on the contrary, by the first special case, ruling out just $A\&\neg B$ is enough for $b(B|A) = 1 = b(X)$. Contradiction.

The principle appealed to in (1) is:

> If C does not entail D (if there are possible situations in which C is true and D is not true), then certainty that C is consistent with less-than-certainty that D.

Here is a putative objection: let D be 'I am certain that C'. C does not entail 'I am certain that C'. But, it might be held, certainty that C is inconsistent with less-than-certainty that I am certain that C. Now, either we do not have infallible access to our own epistemic states, or we do. If we don't, we have no counterexample: being less-than-certain that I'm certain is not incompatible with being certain. If we do, we restrict the principle to beliefs about whose truth uncertainty is possible. Uncertainty about conditionals is possible, so my use of the principle survives.

More generally: the only possible source of trouble for the principle, as far as I can see, will come from beliefs about one's own epistemic state (trouble akin to Moore's paradox: p and I don't believe that p). Provided that conditionals about matches, kangaroos, Ann's and Bill's whereabouts, etc. are not propositions about the believer's mental state, the use of the principle stands.

If we accept this principle, the above argument throws some light on the puzzle which arose at the end of Section 2, p. 138. Two prima facie desirable properties of indicative conditional judgements:

(i) Minimal certainty that $A \vee B$ (ruling out just $\neg A\&\neg B$) is enough for certainty that if $\neg A, B$; changing the negation sign, minimal certainty that $\neg A \vee B$ (ruling out just $A\&\neg B$) is enough for certainty that if A, B.

(ii) It is not necessarily irrational to disbelieve A yet disbelieve that if A, B.

The truth-functional account satisfies (i) but not (ii). Stronger truth conditions may satisfy (ii), if they allow that the conditional may be false when A is false. But they

cannot satisfy (i): for any stronger truth condition, ruling out just $A\&\neg B$ leaves open the possibility that 'If A, B' is not true. The Thesis satisfies both (i) and (ii): (i) ruling out just $A\&\neg B$ makes $b(B|A) = 1$. Yet (ii) it is possible to have $b(\neg A)$ high yet $b(B|A)$ low. So it is incompatible with both truth-functional truth conditions, and stronger-than-truth-functional truth conditions.

We may generalize. Take any proposition. Either it is entailed by $\neg(A\&\neg B)$, or it is not. If it is, it will satisfy (i) but not (ii) (when substituted for 'if A, B'). If it is not, it may satisfy (ii), but cannot satisfy (i). Conditional judgements interpreted according to the Thesis satisfy both (i) and (ii). So they cannot be interpreted as belief in any proposition.

How does the Thesis achieve what belief in no proposition can? Well, suppose $A\&\neg B$ has been ruled out. This is enough for certainty that B given A, *not* because some proposition or other is true whenever $A\&\neg B$ is false; but because B is true in all the worlds that concern the question whether B if A — the A-worlds. What goes on in the $\neg A$-worlds has nothing whatever to do with thoughts about how likely it is that B given A. A high degree of belief in $\neg A$ is consistent with a low degree of belief in B given A, *not* because some proposition is false in some $\neg A$-worlds; but because the fact that $b(\neg A)$ is high has no bearing at all on whether most, or the most probable, A-worlds are B-worlds.

6.6

Anyone interested in the concept of truth should take note of this result. It is an empirical question how well the Thesis fits our practice of assessing conditionals, and it is a deeper question whether, and if so why, it is a good practice if it does. But — to say the least — there could be people who use 'if' this way. The result tells us that they do not use 'if' to express propositions, evaluable in terms of truth. A previously unnoticed test for the applicability of the concept of truth has presented itself: if judgements of a given type are subject to uncertainty, do uncertain judgements of this type fit the structure appropriate to uncertainty about truth bearers? Someone may object that this whole theory of uncertainty, based on the structure of probability, is wrong. Then it is incumbent upon him or her to give an alternative theory of the logic of uncertainty. It is too important a phenomenon — as it applies to conditionals, and as it applies to other judgements — to ignore.

7 IS TRUTH NECESSARY?

7.1 Compounds

Could it be that the mistake philosophers have made, in trying to understand conditionals, is to treat them as part of fact-stating discourse — as representing the world as being a certain way — and that this is not their function? If so, pressing questions arise. What are we doing when we say or think that if A, B, if not saying

or thinking that this is how things are? What do we aim at, if not to state or think the truth? What is it to be right or wrong in so saying or thinking? How are we to understand the role of conditionals in arguments, if not in terms of preserving truth? And there is the question Lewis raises:

> I have no conclusive objection to the hypothesis that indicative conditionals are non-truth-valued sentences... I have an inconclusive objection, however: the hypothesis requires too much of a fresh start. It burdens us with too much work still to be done, and wastes too much that has been done already... [W]hat about compound sentences that have conditionals as constituents? We think we know how the truth conditions for compound sentences of various kinds are determined by the truth conditions of constituent subsentences, but this knowledge would be useless if any of these subsentences lacked truth conditions. Either we need new semantic rules for many familiar connectives and operators when applied to indicative conditionals... or else we need to explain away all seeming examples of compound sentences with conditional constituents. [1976, pp. 141–142].

Too much of the ship would need rebuilding, says Lewis. However, the particular plank on which he rests his case is far from sound. We do think we know how the truth conditions of compound sentences of various kinds are determined by the truth conditions of constituent subsentences. But this knowledge *is* useless when it comes to conditional subsentences. We do not have a satisfactory general account of sentences with conditional constituents. This may be because we have not yet figured out the truth conditions of conditionals. Or it may be because they don't have any.

First, the truth-functional account[45] gives bizarre results for compounds of conditionals. For example,

> Either, if the Queen is at home she is worrying about me, or, if the Queen is not at home she is worrying about me

is a tautology; so, if I reject the first disjunct, I must (on this account) accept the second. And the following argument is valid if we treat 'if' truth-functionally:

> If God does not exist, then it's not the case that if I pray my prayers will be answered. I do not pray. Therefore God exists.

Second, the attempts by Grice and Jackson to explain away the seemingly paradoxical features of truth-functional conditionals, focus exclusively on what more

[45] Lewis holds that indicative conditionals have truth-functional truth conditions, and accepts Jackson's account of their assertability conditions. See the Postscript to 'Probabilities of Conditionals and Conditional Probabilities' [Lewis, 1986, pp. 152–156].

is needed to justify the *assertion* of 'If A, B', beyond the belief that the truth condition is satisfied. They are silent about the occurrence of conditionals, unasserted, as constituents of longer sentences. Jackson is explicit on this point. 'It simply doesn't follow from the fact that I give $(A \rightarrow B)$ truth conditions that I must find, say, $[(A \rightarrow B) \rightarrow C]$. . . a meaningful sentence' [1987, p. 129]. That is, it is compatible with his account of conditionals as (a) having truth-functional truth conditions and (b) being subject to a special rule of assertability, that unasserted conditionals are meaningless. If we want to give them meaning, we have more work to do.

Third, non-truth-functional truth conditions also have controversial consequences for compounds of conditionals. Stalnaker adopts as a logical truth the Law of Conditional Excluded Middle: (if A, B) or (if A, $\neg B$). Lewis admits that there is much to be said for this — he calls it '[t]he principal virtue and the principal vice of Stalnaker's theory' [Lewis, 1973, p. 79] but thinks there is more to be said against it.[46] Another controversy is whether 'If A, then if B then C' is equivalent to 'If A and B, then C'. We do treat these forms as interchangeable, it seems. But on Stalnaker's semantics (or Lewis's for counterfactuals) neither entails the other. Consider (1) 'If it rains or snows tomorrow, and it doesn't rain tomorrow, it will snow tomorrow'. That, it is agreed, is unassailable. Now consider (2) 'If it rains or snows tomorrow, then if it doesn't rain tomorrow, it will snow'. We read that in the same way — just as trivial. But on Stalnaker's semantics (2) may well be false. If snow is a far-out possibility, and rain a close-in possibility, then in all the closest worlds in which it rains or snows, it rains but doesn't snow. Then, the closest world in which it rains or snows (viz. rains) may be such that the closest world to *it* in which it doesn't rain, it doesn't snow either. So for Stalnaker 'If it rains or snows tomorrow, then if it doesn't rain, it won't snow' may be true.

This is somewhat counterintuitive. However, maintaining the equivalence of (1) and (2) also exacts a price: modus ponens for conditionals with conditional consequents. I accept, as trivial, 'If it rains or snows, then if it doesn't rain, it will snow'. I accept that it will rain or snow (because I am nearly certain that it will rain). But I deny that if it doesn't rain it will snow (because I'm virtually certain that if it doesn't rain, it won't snow either).[47]

Turning from particular theories to the phenomena themselves, let's first consider disjunctions of conditionals. 'Or' is a very useful word, especially when it connects things we can be uncertain about, for often we can be confident that A or B, while not knowing which. We can be uncertain about conditionals. Yet 'Either (if A, B) or (if C, D) — but I don't know which' is a form of thought that is virtually uninstantiated. An agile mind will leap to the challenge and instantiate it —

[46] Lewis's remark is about the tenability of this law for counterfactuals.

Stalnaker does not think that there always must be a closest A-world. When B is true in some but not all of the closest, he holds that each disjunct is indeterminate but the disjunction determinately true. The analogue is the treatment of vague terms such that an object may be not determinately red, nor determinately orange, but determinately either red or orange [Stalnaker, 1981].

[47] Examples like this are the topic of Vann McGee's 'A Counterexample to Modus Ponens' [McGee, 1985]. The phenomenon is mentioned by Adams [1975, p. 33].

but once you've seen the results, you see why we have no use for such thoughts. 'Either if I go out I'll get wet, or if I turn the television on I'll see cricket (I don't know which).' That's not too hard to interpret: if it's raining and I go out I'll get wet; if it's not raining and I turn the television on, I'll see cricket; and of course, it's either raining or it isn't. (The disjunction of conditionals has disappeared.) Others need more background: 'Either, if you open box *A*, you'll get ten pounds, or, if you open box *B*, you'll get a button, I don't know which.' If Fred is in a good mood he has put ten pounds in box *A* and twenty pounds in box *B*. If Fred is not in a good mood he has put a paper clip in box *A* and a button in box *B* ... . Again, the disjunction of conditionals is an exceedingly bad way to convey the information you have, and once the necessary background is filled in the disjunction belongs elsewhere. On the other hand, our genuine need for disjunctions shows up naturally inside a conditional: 'If *A*, then either *B* or *C* (I don't know which)'. Some apparent disjunctions of conditionals are no such thing: 'Either we'll have fish, if John arrives, or we'll have left-overs, if he doesn't'.

Turn to negations. If someone makes a remark, e.g. 'It will rain', you may disagree in two ways, one stronger than the other. You may say 'No it won't'; or you may say 'I wouldn't be so sure'. In the first case, you assert the negation of the first statement; in the second, you are prepared to assert neither it nor its negation. Similarly, if someone says 'If it rains, they will be delayed', you may disagree in two ways. If you disagree strongly, you will say 'No, if it rains, they won't be delayed'. Or again, you may go less far, and express uncertainty about whether they will be delayed if it rains. If the analogy holds, then *A* is to $\neg A$ as 'If *A*, *B*' is to 'If *A*, $\neg B$'. And 'It's not the case that if *A*, *B*' has no clear established sense distinguishable from this.

Conditionals in antecedents of other conditionals are also problematic. Gibbard suggests [Gibbard, 1981, pp. 234–238] that we have no general way of decoding them, and some cannot be deciphered; for example 'If Kripke was there if Strawson was, then Anscombe was there'. If someone utters a sentence of this form, we do our best to interpret it by ad hoc strategies. For instance, we can sometimes identify, in context, the obvious basis, *D*, for an assertion of 'If *A*, *B*', and interpret 'If (*B* if *A*), then *C*' as 'If *D* then *C*':'If the light will go on if you press the switch, the electrician has called' (If the power is on, the electrician has called). Michael Dummett ([1973, pp. 351–354]; see also [1992, pp. 171–172]), suggests that some may be understood as saying 'If you accept that *B* if *A*, you must surely accept this': 'If John should be punished if he took the money, then Mary should be punished if she took the money'.

Consider the schema,

> 'If *A*, *B*' is true if and only if, if *A*, *B*.

It makes two claims:

> If (if *A*, *B*), then 'If *A*, *B*' is true.
>
> If not (if *A*, *B*), then 'If *A*, *B*' is not true.

If negation of conditionals, and conditional antecedents, are ill-understood, so is the schema,[48] as Dummett comments [Dummett, 1992, p. 171]: 'we do not know how to interpret this, because it is not our normal practice to apply negation to an entire conditional statement'; and 'we have hardly any use, in natural language, for conditional sentences ... in which the antecedent is itself a conditional, and hence we cannot grasp the content of the principle'.

Conditionals do not go into truth-functional contexts, or into each other, easily, then. (Appiah [1985, pp. 205–210] argues likewise.) Those we do understand, e.g. conditionals in consequents, we understand as equivalent to sentences without embedded conditionals. The facts square at least as well with the hypothesis that conditionals do not have truth values as with the hypothesis that they do. (In Section 9.4 I examine some creative attempts to develop a language with compounds of conditionals which satisfy the Thesis.)

7.2 Validity

Turn to the question of the validity of arguments which involve conditionals. Another reason for disinclination to rebuild the ship might be put: 'Validity is the necessary preservation of truth. Conditionals occur in valid arguments. So conditionals must have truth values'. This conception of validity may be too narrow, independently of conditionals. There are valid arguments involving moral judgements, but it is controversial whether moral judgements have truth values. Legal experts spend their lives deriving consequences from laws, yet it's not obvious that laws have truth values.

Adams [1966; 1975, Ch. 2], gives an account of validity for arguments involving conditionals which conform to the Thesis. His method is far from ad hoc: it teaches us something about classical validity too. He shows that classical, truth-preserving valid arguments are, in a special sense to be made precise, probability-preserving. And this property can be generalized to apply to arguments with conditionals. The valid ones are those which, in the required sense, preserve probability or conditional probability.

Begin with valid arguments which don't contain conditionals. We use them in arguing from contingent premisses which are often believed with less than certainty. The question arises: how certain can we be of the conclusion of the argument, given that we think, but are not sure, that the premisses are true? Call the uncertainty of a proposition one minus its probability. Adams shows this: if (and only if) the argument is valid, then in no probability distribution does the uncertainty of the conclusion exceed the sum of the uncertainties of the premisses. Thus, if I have a valid argument with two premisses each at least 99% probable, this guarantees that the conclusion is at least 98% probable. In this sense, valid arguments are probability-preserving. (They are not probability preserving in a

[48] Wright [1992, pp. 12–20] argues in this way (not in connection with conditionals) that the schema has more substance than might appear at first blush.

different, stronger sense: the probability of the conclusion can be less than the probability of each individual premiss. The Lottery Paradox shows this vividly. We can't expect that much. The conclusion can inherit a risk of falsehood from each premiss, and hence be less probable than each. Still, Adams' result vindicates deductive reasoning from uncertain premisses, provided they are not too uncertain and there are not too many of them.)

This is an independently useful and important consequence of classical validity, then. Now Adams extends this idea to arguments containing conditionals. Take a language with 'and', 'or', 'not', and 'if' — but with 'if' occurring only as the main connective in a sentence. (Thus we put aside compounds of conditionals.) Take any argument formulated in this language. Consider any probability distribution over the sentences in the argument which assigns non-zero probability to the antecedents of all conditionals, that is, any assignment of numbers to the non-conditional sentences which conforms to the Partition Principle, and an assignment of numbers to the conditional sentences which conforms to the Thesis: $p(B \text{ if } A) = p(A\&B)/p(A)$. Extend the term 'uncertainty' to cover conditional uncertainty: the uncertainty of 'If A, B' is one minus $p(A\&B)/p(A)$. Define a valid argument as one such that there is no probability function in which the uncertainty of the conclusion exceeds the sum of the uncertainties of the premisses. And a nice logic emerges — the same as that given by Stalnaker [1968], restricted to simple conditionals. For example, if $p(A) = 0.9$ and $p(B \text{ if } A) = 0.9$, we can show that the lower limit for $p(B)$ is 0.81 (modus ponens has a slightly higher lower limit for the conclusion than can be guaranteed in general).

We saw above (p. 168) that in all distributions, $p(A \supset B) \geq p(B|A)$. Take an argument with conditionals among the premisses but a non-conditional conclusion. Suppose it is valid if we interpret 'if' truth functionally. Then it is also valid in this probabilistic sense — if the conclusion follows from the weaker $A \supset B$ premiss, it follows from the stronger 'If A, B'. But not all truth-functionally valid arguments with conditional conclusions remain valid: the premisses may entail the weaker $A \supset B$ yet not entail 'If A, B'.

Conditional Proof fails. For example, (1) '$\neg(A\&B)$; A; so $\neg B$' is valid, but (2) '$\neg(A\&B)$; so, if A, $\neg B$' is not.

Probabilities can be modelled by proportions, and I shall use them to illustrate the structure behind these facts. (1) If almost everything is A, and almost nothing is $A\&B$, it follows that almost nothing is B. Indeed, if 99% of the things in question are A, and only 1% are $A\&B$, so that 99% are $\neg(A\&B)$, it follows that at most 2% are B, at least 98% are $\neg B$. That's the structure behind the validity of the first case. (2) Suppose that 99% of the things are neither A nor B, and the remaining 1% are $A\&B$. Thus, 99% are $\neg(A\&B)$ but every A is B — 0% of the As are $\neg B$. That's the structure behind the invalidity of the second case.[49] (Ann and Bill are inseparable.

[49] Proportions provide just a model — a structural isomorphism. If you imagine the space of possibilities divided up into enough equally probable little bits, or 'worlds', you can translate the model to 'almost all worlds are A-worlds', etc.

I can believe that it's not the case that Ann-and-Bill are there, without believing that if Ann is there, Bill isn't.)

All the departures from truth-functional validity can be traced to the failure of conditional proof. In the following list, the inference on the left is valid, its partner on the right, derivable by a step of conditional proof, is not.

	Valid	Invalid
(l)	$A;B \vdash A$	$A \vdash \text{If } B, A$
(2)	$A \vee B;\ \neg A \vdash B$	$A \vee B \vdash \text{If } \neg A, B$
(3)	$\neg(A \& B);\ A \vdash \neg B$	$\neg(A \& B) \vdash \text{If } A, \neg B$
(4)	$\text{If } A, B;\ \text{if } B, C;\ A \vdash C$	$\text{If } A, B;\ \text{if } B, C \vdash \text{If } A, C$
(5)	$\text{If } A, B;\ \neg B \vdash \neg A$	$\text{If } A, B \vdash \text{If } \neg B, \neg A$

Models like the above will show this.

Here is one reason why the arguments on the right seem valid: if you are 100% certain of the premisses — if you give them probability or conditional probability 1 — you must give the conclusion probability or conditional probability 1. The counterexamples depend crucially upon at least one premiss being, however slightly, less than certain. Where uncertain premisses are not at issue — in mathematics, say — these inference forms won't let us down. (Analogy: if all *A*s are *B* and all *B*s are *C*, then all *A*s are *C*; but we can have all *A*s are *B*, almost all *B*s are *C*, yet all *A*s are ¬*C*.) We *could* call an argument 'valid' if it satisfied this Certainty criterion, and thus reinstate the arguments on the right. But, in arguing about contingent matters, 100% certainty for our premisses is rare; moreover, it is hard to distinguish from its near neighbours. Knowing that an argument is 'valid' in this sense would be of little use: it would guarantee nothing about what we should think about the conclusion when our premisses are only a hair's breadth away from certainty.

7.3 Speech Acts

'What am I doing, when I say, or think, that if *A*, *B*, if not saying (thinking) that something is the case?' As far as thinking goes, the idea of believing that *B* under the supposition that *A*, of having a conditional belief, of believing something given a hypothesis — Ramsey's idea — is, I hope, clear enough. Here I focus on saying.

Someone asks me who will win the Boat Race. I say 'Oxford will win'. I express a belief. But I speak about the world. If I say 'If the water is calm, Oxford will win', I express a conditional belief; but it is implausible that that is all I do: I also speak about the world — about the Boat Race — albeit conditionally. The answer which fits the Thesis best is this: I make a conditional assertion. My high degree of belief that Oxford will win if the water is calm, amounts to thinking (the water is calm and Oxford will win) is much more likely than (the water is calm and Oxford won't win). I take myself to be in a position (ceteris paribus) to assert that

Oxford will win, not categorically, but conditionally upon the water being calm.[50]

Any kind of speech act can be performed unconditionally or conditionally. There are conditional questions, commands, promises, agreements, offers, etc., as well as conditional assertions. Any kind of propositional attitude can occur within the scope of a supposition. There are conditional beliefs, desires, hopes, fears, etc. 'If he phones, what shall I say?'; 'If he phones, hang up immediately'; 'I want to speak to him if he phones'; 'If he phones, I hope you won't be rude'. It is overwhelmingly plausible that the clause, 'if he phones', does the same job in conditional statements, commands, questions, promises, expressions of wish, etc.; and hence that a theory of conditionals should be applicable to more than conditional statements.

This is quite a severe test. Try applying Stalnaker's theory to conditional commands. Interpret 'If it rains, take your umbrella' as 'In the closest possible world in which it rains, take your umbrella'. Now suppose I have forgotten your command or alternatively am inclined to disregard it. However, it doesn't rain. In the closest worlds in which it does rain, though, I don't take my umbrella. So, on Stalnaker's analysis, I have disobeyed you. Similarly for conditional promises: on this analysis, I could break my promise to go to the doctor if the pain gets worse, even if the pain gets better. This is wrong: conditional commands and promises are not requirements on my behaviour in other possible worlds.

We have already seen that conditional belief is not belief in the truth functional conditional. Nor are conditional commands or expressions of desire, commands etc. that the truth-functional conditional be true. 'If you write the article, submit it to *Mind*.' Now 'Either you won't write the article, or submit it to *Mind*' is a non-starter, not even grammatical.[51] Construed as a command to make the truth-functional conditional true, it amounts to the command 'Either don't write the article, or submit it to *Mind*'. But I am not urging that: you could easily make

[50]Ceteris paribus. Some people interpret the Thesis as an account of when a conditional sentence is assertable: a conditional is assertable to the extent that $b(B \text{ given } A)$ is high. Adams did in his early writings, but not in his book [Adams, 1975] or subsequently. Appiah [1985] does, as do Lewis and Jackson (see Section 9.1). I do not. I interpret it as an account of belief that B if A, to various degrees. Firstly, whether a sentence is assertable depends on all sorts of Gricean contextual factors, which have to be put aside for an account in terms of assertability. Secondly, both for an unconditional claim (Oxford will win) and a conditional one (they'll win if the water is calm), how high $b(B)$ or $b(B|A)$ has to be for an assertion unqualified by 'probably', or 'I think', is a context-dependent matter. Thirdly, this may be a question not just of how close to certain one is, but of the nature and prominence of this uncertainty. Dudman [1992] says we don't assert 'I won't win the lottery' or 'if I buy a ticket I won't win', even if our chance of winning is one in fifty million. Lowe [1995] says likewise. I'm not sure whether they are right, but if they are this has no bearing on the Thesis as I understand it, the claim that someone who knows the chance has a high degree of belief that she won't win/won't win if she buys a ticket.

If no uncertainty is compatible with unqualified assertion we should assert very little. We do assert, unqualified, many things, conditionally or otherwise, where there is more than a chance of one in fifty million that our expectations will be thwarted.

[51]The grammatical fact that conditionals have a main and a subordinate clause fits the view that they are used to do whatever the main clause does, but conditionally.

that true in ways which would please me least of all. Turn the example into an expression of desire: 'If you write the article (W), I want you to submit it to *Mind* (S)'. The conditional desire amounts to a preference for $W\&S$ over $W\&\neg S$. It does not amount to a preference for $\neg W \vee S$ over $W\&\neg S$.[52] For although $W\&\neg S$ is less desirable than $W\&S$,it may be very much more desirable than $\neg W$, which, alas, is a far from implausible way that $\neg W \vee S$ could be true.

The claim that we use if-sentences to make conditional assertions is made by von Wright [1957, p. 131], and is mentioned by Quine [1952]. Quine says:

> Now under what circumstances is a conditional true? Even to raise this question is to depart from everyday attitudes. An affirmation of the form 'if p then q' is commonly felt less as an affirmation of a conditional than as a conditional affirmation of the consequent. If, after we have made such an affirmation, the antecedent turns out true, then we consider ourselves committed to the consequent, and are ready to acknowledge error if it proves false. If on the other hand the antecedent turns out to have been false, our conditional affirmation is as if it had never been made. [1952, p. 19]

As it stands, this last sentence is absurd. It is not absurd if we delete the word 'conditional' from it. It is not absurd to hold that I do not count as having made an assertion unless the antecedent is true. But it is absurd to say it is as if I had not made a conditional assertion — as if I had said nothing at all. I say to you 'If you press that switch, there will be an explosion'. As a consequence, you don't press it. Had I said nothing at all, let us suppose you would have pressed it. A disaster is avoided, as a result of this piece of linguistic communication. It is not as if nothing had been said. This is no objection to the idea that I did not (categorically) assert anything. For let us suppose that I am understood as having made a conditional assertion of the consequent. My hearer understands that if she presses it, my assertion of the consequent has categorical force; and, given that she takes me to be trustworthy and reliable, if it does acquire categorical force, it is much more likely to be true than false. So she too acquires reason to think that there will be an explosion if she presses it, and hence a reason not to press it.

Dummett, like Quine, misrepresents the notion of a conditional assertion when he says it is 'as if [someone] had handed his hearers a sealed envelope marked 'Open only in the event that...' .' [Dummett, 1992, p. 115]. If it were like that, modus tollens would be impossible, as Dummett points out. Whereas, on the lines of the example above, we can explain why someone infers that $\neg A$ when he knows that B is false and a trustworthy person has just asserted B conditionally upon A. Elsewhere [Dummett, 1973, p. 341ff], Dummett is sensitive to the difference between 'no (categorical) assertion has been made' and 'nothing has been said'. It' s not just that the sealed-envelope interpretation cannot be true of our use of conditionals. There is something intrinsically absurd in the idea that understanding

[52] That is, it does not entail this preference; although it is compatible with it.

a sentence should require you, in certain circumstances, to behave as though it had not been said. 'We cannot lay down a convention that no one is to be influenced' (op. cit. p. 342). I undertake to care for your children if you die. Even if you don't die, my conditional undertaking has consequences for you and for me.

It is the analogue of the mistaken, sealed-envelope interpretation for conditional commands that leads Dummett to say that a conditional imperative where the antecedent is in the agent's power 'must... be interpreted as a command and to make the material conditional true' [Dummett, 1973, p. 340]; see also [1959, pp. 8–9]. A child is told 'If you go out, wear your coat'. If he cannot find his coat, he stays in, in order to comply with the command. On my interpretation, if the child can't find his coat, he has a choice between disobeying the command, and behaving in such a way that no categorical command has been made (not: behaving as though nothing had been said). If he wishes not to disobey, he must stay in. Dummett claims that there is no distinction between not disobeying a conditional command, and obeying it. But other examples make this implausible. If, in the emergency ward, you're told 'If the patient is still alive in the morning, change the drip', and you smother the patient, you can hardly claim to have merely carried out an order.

A conditional assertion 'If A, B' is an assertion of B when A is true, and an assertion of nothing when A is false. It is natural then, to say my conditional assertion is true if A and B are both true, and false if A is true and B is not, and has no truth value when A is false. This is compatible with the Thesis, provided we interpret this assignment of truth values with care. Belief that if A, B is not belief that it is true. For it is true only if $A\&B$, and we may believe that if A, B without believing that $A\&B$. Nor is it belief that it is not false. For it is not false provided that $\neg(A\&\neg B)$, i.e. $A \supset B$; and we can believe that it is not false without believing that if A, B. Belief that if A, B is a conditional belief that it is true given that it has a truth value — belief that it is true given that it is either true or false. Now my degree of belief that 'If A, B' is true, given that it has a truth value, is just $b(A\&B)/b(A)$, as it should be. (The bombshell is avoided because belief that if A, B is not belief that something is true.) For a proposition assumed to be either true or false, your degree of belief that it is true given that it has a truth value, is your degree of belief that it is true. So our proposal is not ad hoc. It has as a special case that for a bivalent proposition, to believe it is to believe that it is true.

In making conditional assertions, we do not aim at truth (for we don't assert them only if we believe $A\&B$); nor do we aim at avoiding falsity (for we don't assert them whenever we believe $\neg(A\&\neg B)$); our aim is that they be true given that they have a truth value — that if it turns out that A, we get B as well, rather than $\neg B$.

Dummett [1959, pp. 10–14] rightly says that giving a truth table for a statement with three values, T, F, and X, gives you little guidance as to how the statement is to be used. Does the speaker intend to rule out X (as in the case of empty names), or not? On the conditional, Dummett says '[the speaker] is not taken as having misused the statement or misled his hearers if he envisages it as a possibility that that case will arise in which he is said not to have made a statement true or false'

(p. 11). That is correct, but it does not follow that for conditionals X is really a species of truth, as Dummett claims (p. 12). There is a difference between the claim that I may use this statement correctly when the antecedent is false, and the claim that whenever the antecedent is false, I have used the statement correctly.

The 'true, false, neither' classification does not yield an interesting 3-valued logic or a promising treatment of compounds of conditionals (for it is not a case of there being some 'designated' value or values). It helps only in minor ways. It allows us to say that a conditional is straightforwardly false if its antecedent is true and the consequent false; and that it is straightforwardly true if the antecedent and consequent are both true. There is nothing comparably straightforward to say when the antecedent is false.

An unconditional assertion, e.g. that John is in London, can be right by good luck, or wrong by bad luck: my reasons can be good yet I'm wrong, my reasons can be bad yet I'm right. Or I can have no reasons, yet guess, or have a hunch that John is in London. Likewise for a conditional assertion, that John is in London if Mary is. Mackie [1973, p. 107] has a father saying to a child 'If you put your finger through the bar [of the monkey's cage], it will be bitten off'. The child does so nevertheless. In one scenario, the monkey pays no attention, but a bird swoops down and bites off the finger. The conditional assertion was true, for an unexpected reason. In another scenario, the monkey is about to bite when a rock falls, squashes the cage and kills the monkey. The assertion was false, for an unexpected reason. If it is plausible that a conditional assertion, like an unconditional one, can be right or wrong by luck, this is an argument against those who insist that the antecedent must be 'relevant' to the consequent (see pp. 34–37). And it adds plausibility to the feature of the Thesis to which such people object: belief that $A\&B$ is sufficient for belief that B if A. If, as you believe, $A\&B$ is true, so is your assertion of B on the condition that A.[53]

8 OBJECTIVITY AND ITS LIMITS

8.1

The truth values permitted by the notion of conditional assertion are little more than epiphenomenal — they don't significantly change the picture we had without them. We need more objectivity for conditionals than they provide, it will be complained: we need an account of how a conditional can be right, or wrong, even if its antecedent is false. 'If that lump of sugar is placed in water, it will dissolve' is *true*, it will be said; 'If that lump of granite is placed in water, it will dissolve' is *false*, even if neither is placed in water. Yet more decisive, 'If it's square, it has

[53] I became more aware of the relevance of other conditional speech acts on reading Michael Firestone's thesis, 'The Meaning of "If" ' (Australian National University). Michael Woods' [1997] also treats 'simple conditionals' as conditional assertions; he too has a careful examination of the different conditional speech acts.

4 sides' and 'If it's square, it has 5 sides' are (surely!) true and false respectively, even if the object in question is not square. Or consider my conditional assertion of 'If you press the button, there will be an explosion'. You don't press it. We hold a post mortem, trying to establish whether there would have been an explosion if you had pressed it (NB). It could end with 'You see, I was right', or with 'You were wrong — there would have been no explosion if I had pressed it'.

From what we have seen so far, the Thesis need not rule out there being an objectively correct thing to think about whether *B* if *A*. The right degrees of belief that it has 4 sides/5 sides given that it's square are 1 and 0 respectively. In the case of the sugar lump and the granite, the right degrees of belief (in normal cases — putting aside super-saturation and the like) are at least very close to 1 and 0. 90% of the red balls in this bag have black spots. You are to shake it, put your hand in, and pick a ball. There is a right degree of belief that the ball you pick will have a black spot if it's red. For the best opinion about whether you'll be cured if you have this operation, ask the best doctor you can find. The chance that the chemical substance will emit dangerous radiation if stored underground, will be best estimated by a chemist or physicist. You read in the newspaper that if you eat garlic, you are less likely to get heart disease. You watch the weather forecast. And so on. We all have the idea of a right, or at least a better, opinion. That is, we have the idea of objective probability — or at least the idea that some degrees of belief are worth more than others. An expert is someone who has acquired good judgement in a given area — and moreover, has access to more relevant information than the rest of us, in that area. An expert is someone whose advice we do well to heed, in forming our own beliefs and plans for action.[54]

Prima facie, there is room for an account of objectively correct conditional thoughts. It doesn't follow that they have truth conditions. The following has been suggested.[55]

> 'If *A*, *B*' is true iff the objective probability of *B* given *A* is sufficiently high.

This is not compatible with the Thesis, and is independently objectionable. (I do not object to the fact that the truth condition is vague.) Presumably, in a context, either there is some number less than 1 which is sufficiently high; or there is some number greater than 0 which is not sufficiently high; or (most likely) both. Take an example where objective probabilities are relatively easy to estimate — balls in bags, say. Call the proposed truth condition *S*. First suppose 0.9, say, is sufficiently high, and I am certain that the objective probability of *B* given *A* is 0.9. My degree of belief in *B* given A is 0.9. According to the Thesis, I am 90% confident that if *A*, *B*. But I am certain that *S*, hence, certain that the conditional is true. By the truth condition, I am 100% confident that if *A*, *B*. (The truth condition has the additional

[54] Adams is no subjectivist about probability. A section of his book [Adams, 1975] is entitled 'A motive for wanting to arrive at correct probability estimates'.

[55] See Simon Blackburn ([Blackburn, 1986], pp. 213-5); Michael Woods [1997]; and the suggestion crops up orally from time to time.

embarrassing consequence that the truth of 'If A, B' is compatible with the truth of $A\&\neg B$.) Second, suppose some number greater than 0 is not sufficiently high — 0.5 say. Suppose that I am certain that the objective probability of B given A is 0.5, and so have degree of belief 0.5 that B given A. By the Thesis, I am 50% confident that if A, B. Now I am certain that S is false, hence certain that the conditional is false. By the truth condition, I am 0% confident that if A, B. (The truth condition also has the consequence that the truth of $A\&B$ is compatible with the certain falsity of 'If A, B'. Not everyone minds that. *I* think it's wrong for me to say 'It is certainly false that, if you approach, the dog will bite', when I know that the objective conditional probability of its biting, given that you approach, is 0.5; and further, to admit no error when you approach and are bitten — to stick to my judgement that the conditional was certainly false. But not everyone agrees with me.[56]

If we are to have objective values, we need values intermediate between truth and falsity. But there is an obstacle to objective values — to there being a right thing to think — to which I now turn.

8.2

Gibbard [1981, pp. 231–232] presented an argument for the Thesis and against truth for indicative conditionals, which threatens to wipe out objectivity along with truth. This is its structure as I see it. (1) If two statements are compatible, so that they can both be true, a person may consistently believe both of them simultaneously. (2) For consistent A, and any B, people do not simultaneously believe both 'If A, B' and 'If A, $\neg B$' (unless by oversight), nor consider it permissible to do so; rather, to accept 'If A, B' is to reject 'If A, $\neg B$'. (This accords with the Thesis: if $b(B|A)$ is high, $b(\neg B|A)$ is low. It also accords with Stalnaker's truth conditions but not with the truth-functional account.) So, by (1) 'If A, B' and 'If A, $\neg B$' can't both be true: if they could, why shouldn't someone readily accept both? But (3) one person X can have impeccable reasons for believing 'If A, B', while another person Y has impeccable reasons for believing 'If A, $\neg B$'; (a) the situation is symmetric: there is no reason to prefer X's belief to Y's, or vice versa; no case can be made for saying just one of the beliefs is false; (b) neither of them is making any sort of mistake; each is rational, and bases his judgement on known truths; no case can be made for saying both beliefs are false. So: they can't both be true, they can't both be false, and it can't be that just one of them is true. Truth and falsity are not suitable terms of assessment for conditionals.

Gibbard's much-discussed example is the Sly Pete Story. It concerns a poker game. Jack saw that Pete had the losing hand, and believes 'If Pete called, he lost'. Zack knows that Pete, the cheat, knew the contents of his opponent's hand, and

[56] Pendlebury [1989], Read [1995] and others argue that the truth of $A\&B$ is compatible with the falsity of 'If A, B'.

that Pete always plays to win. He believes 'If Pete called, he won'. In fact, Pete didn't call. Once they learn this, neither has any use for a thought beginning 'If Pete called'.

Gibbard's example is perhaps not perfectly symmetric, and some have argued that Jack's belief is better than Zack's. Pendlebury [1989, p. 182] claims that from the God's eye point of view, Jack's conditional is the true one. This points, at best, to an imperfection of the example. Here is a boring, perfectly symmetric one. In a game, (1) all red square cards are worth 10 points, and (2) all large square cards are worth nothing. *X* caught a glimpse as *Z* picked a card and saw that it was red. Knowing (1), he believes 'If *Z* picked a square card, it's worth 10 points'. *Y*, seeing it bulging under *Z*'s jacket, where *Z* is keeping it out of view, knows it's large. Knowing (2), he believes 'If *Z* picked a square card, it's worth nothing. (Someone who knows all the relevant facts knows it isn't square, and has no use for a conditional beginning 'If it's square'.)

There is little hope for objectively correct opinion, if one person can have a completely adequate reason to accept 'If *A*, *B*', and reject 'If *A*, ¬*B*', while another has a completely adequate reason to do precisely the opposite. It is not as though, if either had more information, he would know what to think. If he had more relevant information, he would know that ¬*A*, and have no use for either conditional, 'each of which is a ticket for an intellectual journey starting at a place where he knows he will never be' [Bennett, 1988, p. 520].

How widespread is the Gibbard phenomenon? Gibbard [1981, pp. 226–229] thinks there are two kinds of conditionals: 'epistemic' ones, which satisfy the Thesis, and are subject to this phenomenon, paradigms of which are past-tense indicatives; and 'nearness conditionals', to be treated à la Stalnaker, which are not subject to the phenomenon, paradigms of which are subjunctive conditionals. Future-tense indicatives can function as either (p. 228). I don't think conditionals divide in this way. For any contingent conditional, the world may be such that the Gibbard phenomenon can arise. Here is another example.

Suppose there are two vaccines against a certain disease, *A* and *B*. Neither is completely effective against the disease. Everyone who has *A* and gets the disease gets a side effect *S*. Everyone who has *B* and gets the disease doesn't get *S*. Having both vaccines is, however, completely effective against the disease (though not many people have both). These scientific facts are known. *X* knows that Jones has had *A*, and says 'If Jones gets the disease, he'll get *S*'. *Y* knows that Jones has had *B*, and says 'If he gets the disease, he won't get *S*'. (If Jones is meanwhile run over by a bus and killed, these can go counterfactual: 'If he had got the disease, he would have got *S*', 'If he had got the disease, he wouldn't have got *S*'. But counterfactuals are to be put aside until Section 10.) In all these cases, if the full story is known, the conditionals become useless. For instance, the doctor giving Jones *B* says 'I can't guarantee that you won't get the disease, but if you do, you won't get *S*'. 'But I've already had *A*', says Jones. 'Oh well, then, you won't get

the disease (the question of what will happen if you do doesn't arise)'.[57]

Now let us change the initial story slightly. Everyone who has just A and gets the disease, gets S. Everyone who has just B and gets the disease, doesn't get S. Few people have both. Of those who do, very few — say 0.01% — get the disease (whereas about 1% of those who have just A, or just B, get the disease). Anyone who has both vaccines and gets the disease has a 50% chance of getting S. These are known facts. As before, X knows that Jones has had A. He has grounds for near-certainty that if Jones gets the disease, he will get S. Y, who knows that Jones has had B, has grounds for near-certainty that if he gets the disease, he won't get S. This time, neither has the right opinion. There is further obtainable information which would lead each to a better opinion: anyone who knows the relevant scientific facts, and that Jones has had both vaccines, thinks that the chance that Jones will get S if he gets the disease is 0.5. Doctor, having given Jones B: 'I can't guarantee that you won't get the disease, but if you do, you won't get S'. Jones: 'But I've already had A'. Doctor: 'Oh, then I must correct what I said: it's very unlikely that you will get the disease; and it's 50% likely that you will get S if you do get the disease'.

The Gibbard phenomenon arises if and only if there are currently ascertainable facts which rule out A.[58] In one direction: let F be a set of currently ascertainable facts from which we can derive that $\neg A$. That is, $A\&F$ is inconsistent. That is, $A\&F$ entails a contradiction. Furthermore, it will generally be the case that if we know F, we can learn that $\neg A$ by assuming A and deriving, from $A\&F$, a contradiction, $B\&\neg B$. Now, F is true, therefore consistent. It is the addition of A that enables us to derive $B\&\neg B$. So there must be two subsets of F (not necessarily disjoint), F_1 and F_2, such that from $A\&F_1$ we can derive B, and from $A\&F_2$ we can derive $\neg B$. So someone who knows just F_1 has adequate reason to believe 'If A, B'; and someone who knows just F_2 has adequate reason to believe 'If A, $\neg B$'.

In the other direction: suppose there is no set of currently ascertainable facts which rule out A: all currently ascertainable facts are consistent with A. Therefore, there are not two subsets F_1 and F_2 such that $A\&F_1$ entails B and $A\&F_2$ entails $\neg B$. The Gibbard phenomenon does not arise. There may be two subsets which render it probable, and improbable, respectively, that B given A. But this is no threat to objectivity. People in these states of information can improve their opinions by learning more. There is no obstacle to there being an objectively right thing to think, based on all the relevant currently available information.

If currently ascertainable facts are sufficient to rule out A, then A has a current objective probability of 0. Hence the present objective probability of B given A is undefined. Hence there is no ideal, objective thing to think. (There may of course

[57] I have given elsewhere a more careful example, which relies just on the gas laws [Edgington, 1991, pp. 206–207]. The examples show that conditionals which are based on 'objective connections in the world' are not immune from the Gibbard phenomenon.

[58] We must include general truths among the ascertainable facts; in the examples we have considered, some were founded on the rules of a game; and there was 'Pete always plays to win'; and the scientific general truths about the vaccine.

be more or less rational ways of assessing the information you have. But the world does not make one judgement best.) A necessary condition for there being an optimum judgement, then, is that the antecedent has a present non-zero chance of being true.[59]

8.3

One way of approaching the idea of correct judgement is via the device of an ideal epistemic perspective. There are different degrees to which we can idealise. Take the extreme: G_1 knows everything — all the facts, past, present and future. Consider the conditional 'If it rains tonight (R), the river will overflow its banks tomorrow (O)'.[60] G_1 knows $R\&O$, or knows $R\&\neg O$, or knows $\neg R$. He knows too much to have any use for indicative conditionals. He can pronounce them trivially true or false in the first two cases. If the third case obtains, the question doesn't arise.

G_1 is a little too ideal for us to relate to. Before turning to G_2, let us look at some features of the concept of objective chance.

The concept of objective chance gets some purchase when, at least apparently, like causes do not have like effects; and moreover, in a class of apparently relevantly identical cases, the proportions of the various sorts of outcome are relatively stable, although these proportions are generated in an apparently random way. These facts about proportions could be brute (as those who interpret objective chance as relative frequency think); or they could be explained as just the sort of pattern of outcomes one would expect if (e.g.) each F has a $p\%$ chance of resulting in a G. The strongest notion of objective chance applies only if we remove 'apparently' from the above: relevantly qualitatively identical states do have different outcomes. Short of that, which different outcome occurs may be arbitrarily sensitive to the exact values of the variables which specify the cause: there may be no degree of similarity short of qualitative identity such that, if two situations are that similar, the outcome will be similar.[61] As we cannot measure to an infinite degree of accuracy, we need to apply the concept of chance. And there will be many more cases where detection of relevant difference is not practicable, and application of the concept of chance is useful. We could call these three cases 'absolute indeterminism', 'quasi-indeterminism' and 'practical indeterminism'; there will be correlative notions of determinism, depending on which they rule out. It will not matter which way you interpret 'indeterminism' and 'determinism' in what follows. They represent different degrees of idealisation of epistemic perspectives.

[59] For present purposes, nothing which is now causally possible has a zero chance of being true. We can give infinitesimal probabilities to hitting a particular point on a dart-board, etc. See Lewis [1980, p. 89], McGee [1994].

[60] This example is Bennett's [1988, p. 521], where he also uses the device of an ideal epistemic perspective.

[61] This gives rise to Chaos Theory. See Gleick [1987]; Smith [1991].

The point of the concept of objective chance is that knowledge of the chances of future events gives us reasons for our expectations. There are laws about chances, which we try to ascertain.

The idea of backward-looking objective chance is as recherché as the idea of backward causation.[62] A past event may *have had* a real chance of not coming about. But the time for the realisation of that chance has passed. The event happened, and has no present chance of not having happened. (There may be an exception: the unobserved wave packet of quantum theory which didn't collapse. Our great difficulty in understanding this phenomenon underlines my point about how we normally think.) Chances change with time, according to the outcomes of intervening chance events. The probability of my winning my bet on 3 heads is $\frac{1}{8}$. After one toss, it has changed to either $\frac{1}{4}$ or 0, for the probability of $\frac{1}{2}$ that the first toss had of landing heads has now 'collapsed' to 1 or 0. A man goes through a maze at constant speed, deciding his path by a random device. At any point, we can calculate the chance that he will be at the centre by noon. This can vary as he makes unlucky or lucky turns, until noon at the latest, when it becomes 1 or 0.[63]

Our reasoning about the past respects this asymmetry in the direction of chance. We try to find the best explanation of what we currently know, and examine hypotheses about past forward-looking chances. The chance *was* high that he would get these symptoms, if he took arsenic; the chance was low that he would get these symptoms if he didn't.

Now let us return to the ideal epistemic perspective. G_2 does not have magical knowledge of the future (or anything else). For a given conditional, e.g. 'If it rains tonight, the river will overflow its banks tomorrow', he knows all he needs to know about the past and the laws of nature. (These are things we aspire to.) Now distinguish two cases: if determinism is true, G_2 is in as good a position as G_1. From his knowledge of the past and the laws, he can infer the future. Again, he knows $R\&O$, or knows $R\&\neg O$, or knows $\neg R$. He has no non-trivial use for indicative conditionals. If determinism is false, however, G_2 can know that R has some chance of occurring, and moreover, can know the present chance of O given R — it may be 1 or 0 or something in-between.

But suppose G_2 knows that the chance of rain is now 0. Then $p(O|R)$ is undefined. There is no objective fact about how likely it is, now, that the river will overflow given that it rains, when it is now causally impossible that it rain. Compare the doctor and the vaccination, above (p. 184). (There may be an objective fact about how likely it is that it would have overflowed if it had rained. More on that later.) Objectivity breaks down when the present chance of the antecedent's being true is 0.

When there is a right thing to think, it is only temporarily right. The chance of a future possibility changes with time; and in due course, flips to 1 or 0, according

[62] I leave open the question whether it is incoherent, or merely very weird.

[63] This is Lewis's delightful example of the 'garden of forking paths' [Lewis, 1980, p. 91]. My discussion of chance owes much to this article.

to what actually happens. Consider 'If it rained on Monday night, the river overflowed its banks on Tuesday afternoon'. Even if I know that the chance of O given R *was* 1, someone else can have just as good reason to say 'If it rained on Monday night, the river didn't overflow on Tuesday afternoon' — either because she saw that the river didn't overflow, or perhaps saw from the state of the river on Tuesday morning, or on Wednesday, that it couldn't overflow (or have overflowed) on Tuesday afternoon. For backward-looking indicatives, there is the better or worse management of uncertainty but no ideal view.

Objectively correct present values remain, then, for future-looking conditionals whose antecedent has a present non-zero chance of being true (there are no currently ascertainable facts to rule it out; it is still causally possible). But a difficulty arises for the project of providing stronger-than-truth-functional truth conditions for such conditionals. People hanker after something along Goodman's or Lewis's lines, some sort of strict conditional: the truth of 'If A, B' requires that the truth of B be guaranteed by the truth of A, together with other facts, given the laws — at least if the antecedent is false. (This qualification is unnecessary for those who deny that $A\&B$ is sufficient for if A, B, for instance Pendlebury [1989], Lowe [1995], and Read [1995]. Others, like Lewis, will accept that the conditional is true if it turns out that $A\&B$, even if it is causally possible that $A\&\neg B$. (Lewis is, of course, only giving a theory of counterfactuals; but he applies the theory widely, for example to the forward-looking conditionals needed for decision theory [Lewis, 1981, pp. 325–335].)

A fair dose of determinism, then, is required for truth — to ensure a connection between A and B, and to ensure that all the initial conditions are in place. But total determinism will mean that, if the antecedent is false, it is now causally impossible, and there is no ideal thing to think. I do not say that happy combinations of determinism and indeterminism are impossible. But it is hard to see what would ground our confidence that many conditionals are true. (No general predilection towards determinism, or indeterminism, would.)

If so much is required for truth, so little is required for falsity, on the above views. Consider 'If you toss the coin ten times, it will land heads at least once'. Add that it is improbable that you will toss it. The conditional is either certainly false (on the stronger alternative) or probably false (on the weaker alternative — it is true only if you do toss it and it lands heads at least once, but it is improbable that you will toss it). Our everyday conditionals run great risks of being certain1y false, or much too probably false. It may be replied that charity requires that we interpret them with a silent 'it will be very probable that' inserted in the consequent: 'If you toss the coin ten times, there will be a very high chance that it will land heads at least once'.[64] Well then, truth, as opposed to falsity, for the plain conditionals we utter, isn't what matters about them. We are happy enough with falsity, provided

[64]The conditionals Lewis uses in the analysis of causation and decision, when the assumption of determinism is dropped, have chances in the consequents [Lewis, 1986, pp. 175–184; 1981, pp. 329–335].

that the chance of consequent given antecedent is sufficiently high. It is rather more straightforward to construe people as doing their best, on the basis of as much relevant information as it is worth their while to acquire, to estimate as accurately as possible how likely it is that B given A.

The concept of objective chance is philosophically puzzling. But ordinary people do understand it, when they read that (e.g.) eating garlic reduces one's chance of heart disease. Those who would explain it away need a surrogate for it. I hope that I have not relied on anything too theoretically contentious in the discussion above.

9 IS TRUTH POSSIBLE AFTER ALL?

9.1

According to Jackson [1979; 1980; 1987] we can explain why the Thesis gives the right *assertability* condition for indicative conditionals, while maintaining that their truth condition is the truth-functional one. The relation between '$A \supset B$' and 'If A, B' is analogous to the relation between 'A and B' and 'A but B', he claims. The latter pair have the same truth conditions; but it is part of the meaning of 'but' that it is used to signal a contrast between the propositions it joins. In the case of 'if', it is part of its meaning that it signals not only that the speaker believes that $A \supset B$, but that this belief is robust with respect to the antecedent: the speaker would not abandon the belief if she were to learn that A. So someone who asserts 'If A, B' must not only have a high degree of belief in $A \supset B$, but must also have a high degree of belief in $A \supset B$ given A. But $b((A \supset B)|A) = b(\neg A \vee B|A) = b(B|A)$. So we assert conditionals when we have a high degree of belief in the consequent given the antecedent.

Jackson claims an explanatory advantage over those who take the Thesis as primitive — the 'no-explanation' theorists, he calls them [Jackson, 1987, p. 55]. 'To have assertability conditions best explained by certain truth conditions *is* to have those truth conditions' [Jackson, 1987, p. 58]. It is a plausible methodological maxim that the value of an explanation depends on its explaining more than the data it accommodates. So let us first ask what else, if anything, the truth conditions explain.

Armed with truth conditions, can we explain the occurrence of conditionals as constituents of longer sentences? We cannot. We know how $A \supset B$ behaves in compound sentences. But Jackson's theory is not that 'If A, B' means the same as $A \supset B$: their truth conditions are the same, but their assertability conditions differ. The theory has no implications for how conditionals behave in contexts in which they are unasserted. In fact he adopts the strategy of explaining away such occurrences [Jackson, 1987, pp. 127–137] along the lines of Section 7.1 above.

Do the truth conditions explain the validity of arguments involving conditionals? A main source of dissatisfaction with the truth-functional conditional is a

clash between our intuitions about validity and the arguments it licenses as valid (see the example on p. 171 above). Jackson claims that our intuitions are at fault here: we confuse preservation of truth with preservation of assertability [Jackson, 1987, pp. 50–51]. These generally stay in line, but, because of the special rule for asserting conditionals, here they come apart. Adams gave an account of validity in terms of preservation of probability or conditional probability (assertability for Jackson), which coincides with an account in terms of preservation of truth for arguments without conditionals. Lewis, agreeing that our intuitions about validity go better with Adams, says 'As to whether 'validity' should be the word for truth- or assertability-preservation, that seems a non-issue if ever there was one' [Lewis, 1986, p. 153]. Lewis, presumably, means not merely that we can use the word 'valid' any way we like, but that either choice would be reasonable. Adams' guiding thought was a pragmatic one. Take an argument with two or three premisses. Would it be useful to classify it as 'valid' when one can be arbitrarily close to certain of the premisses yet reject the conclusion utterly (or in Jackson's terms, if the premisses are very highly assertable and the conclusion completely unassertable)? Adams thought not.

Jackson's account does have the advantage that if A is true and B is false, 'If A, B' is straightforwardly false. A defender of the Thesis can point out that a high $b(B|\, A)$ commits you to a high $b(A \supset B)$, and so commits you to something false if A is true and B is false. But this is somewhat indirect. However, in Section 7.3 I argued that one could, compatibly with the Thesis, interpret asserting 'If A, B' as making a conditional assertion, true if $A\&B$, false if $A\&\neg B$, truth-value-less otherwise (see above pp. 179–180).

We have yet to see any explanatory advantage of Jackson's theory. I turn now to some disadvantages. In Section 7.3 I claimed that the notion of a conditional assertion was part of a uniform theory of conditional speech acts — conditional commands, etc. A theory of conditional statements, I claimed, should allow that an if-clause, 'If he phones', plays the same role in 'If he phones, Mary will be pleased' and 'If he phones, hang up immediately' (p. 169–169). This was a difficulty for Stalnaker's theory, and for the truth-functional theory, I argued. Let us try to extend Jackson's theory to conditional commands: to command that if A, B, is not only to command that $(A \supset B)$, but also to signal that you would still command $(A \supset B)$ if you believed that A were true. Return to the example 'If the patient is alive in the morning, change the drip' (p. 179). On this analysis I command 'Make it the case that either the patient is not alive in the morning, or you change the drip'. You obey my command if you kill the patient. There is no obvious reason why you should concern yourself with what I would have commanded had I believed that the patient would be still alive. (This is quite distinct from the notion of a conditional command — an utterance which has the force of a command to make the consequent true, on the condition that the antecedent is true.)

I turn to a difficulty for Jackson's explanation of why we assert a conditional when $b(B|A)$ is high, noticed by Lewis. Having a high degree of belief in B given A (call this, following Lewis, robustness$_1$) does not always mean that you would (or

even think you would) be confident in B if you learned A (call this robustness$_2$). Consider again 'If Reagan is in the pay of the KGB, I'll never find out' (see above, p. 161). I do have a high conditional degree of belief in consequent given antecedent. I do assert the conditional. But I know in advance that I won't believe the consequent if I learn the antecedent. Lewis says

> What really matters is robustness$_2$, so it would be more useful to signal that. On the other hand it would be much easier to signal robustness$_1$... It may be no easy thing to judge what would be learned if [A] were learned, in view of the variety of ways in which something might be learned. For the most part, robustness$_1$ is a reasonable guide to the robustness$_2$ that really matters. So it is unsurprising that what we have the means to signal is the former rather than the latter. And if this gets conventionalized, it should be unsurprising to find that we signal robustness$_1$ even when that clearly diverges from robustness$_2$. That is exactly what happens... I say 'If Reagan works for the KGB, I'll never believe it'. [Lewis, 1986, pp. 155–156].

So: when we utter a conditional 'If A, B', we convey that we believe $A \supset B$; and we would like to convey, in addition, that if we were to learn A, we would still accept $A \supset B$. But that is a hard thing to be confident about, as Lewis admits.[65] We settle instead for something easier: our conditional degree of belief in B given A — not because it is intrinsically interesting in itself, but because it is a good but fallible guide to what we would believe if we learned A. As this convention is established, even in cases where $b(B|A)$ is high but we would not believe B if we learned A, we assert 'If A, B'.

My point about conditional commands, combined with this difficulty, might make us wonder whether it is robustness$_2$ that really matters. With conditional commands, I need not concern myself with what you would command if you learned something else; I need only concern myself with what you do command, albeit conditionally. Similarly with conditional assertions, one might wonder why I should be interested in what you would assert if you learned that A — after all, as has been pointed out, this is a difficult thing to be confident about. I am interested in the fact that you are confident that B on the supposition that A — confident enough to assert that B, conditionally upon the truth of A. Is Jackson's theory of the meaning of 'if' plausible? His favourite analogy is with the meaning of 'but' ([Jackson, 1987], p. 26). Consulting the dictionary on 'but', I see 'in contrast'. On 'if', I see 'on condition that; provided that; supposing that'; I do not find anything to suggest that 'if' conventionally means anything about what I would assert if I learned something.

[65] This, Jackson insists [1987, p. 33], and Lewis no doubt agrees, is a subjunctive (counterfactual) conditional. Jackson's theory of these is similar to Lewis's. And it certainly is hard to be sure that in *all* close worlds in which I learn A, I will accept B. For there might be, for all I know, unexpected ways of learning A — like learning that the match was struck, but at the bottom of the swimming pool.

A thesis like Jackson's — belief that the truth-conditions of a statement *S* are fulfilled is not sufficient for it to be asserted — might be tested by trying to find out the conditions under which people believe that *S*, even when they would not be prepared to assert it. Take a co-operative subject in a context where it is clear that, for purposes of research, we want to elicit her beliefs. First consider 'but'. She believes that Ann is poor, and that Ann is honest. She sees no reason to contrast these states of affairs. She wouldn't *say* 'She's poor but she's honest' (except in a special context where she's hunting for an impecunious crook). She is asked whether she assents to that sentence. She might hesitate. She might say 'Yes — but I wouldn't put it that way'; she might say 'She's poor, she's honest, but why the 'but'? Now she's asked whether she assents to 'If the Tories win, they will nationalise the motor industry'. She thinks the Tories have a very small chance of winning. Still, she unhesitatingly says 'No'. (She might add 'That's certainly false'.)

Jackson is well aware that it is impossible to support his theory by eliciting evidence about when people believe conditionals. He advocates an error theory here — people are wrong about 'if'. We speak and think as though there were a conditional connective '$*$', such that $b(A * B) = b(B|A)$. Not many people know Lewis's 1976 result that there isn't [Jackson, 1987, p. 39]. (It is not surprising, then, that dictionaries do not give the true meaning of 'if'.)

Now in the sense (if any) that ordinary people can be credited with believing that there is a conditional 'connective', there is a conditional connective: take two suitable sentences; make, if necessary, some grammatical changes, add an 'if' in an appropriate place and you have one usable sentence. It is philosophers, not ordinary people, who have misconstrued it.

What about the response 'That's certainly false' to the conditional about the Tories? Doesn't a defender of the Thesis have to attribute error to the speaker? I don't think so: If you say to her 'You mean it's certainly false that the Tories will nationalise the motor industry, on the assumption that they win the next election?', I think she would accept that paraphrase. In Jackson's view, if we were free from error — if we stopped being flat-earthers [Jackson, 1987, p. 40] — we would see that that conditional is very likely to be true when it's very likely that the Tories won't win. In fact, however, we are better off in 'error'. As I said earlier (p. 135) we would be intellectually disabled without the ability to discriminate between believable and unbelievable conditionals whose antecedent we think is false.

9.2

Mellor defends a position in some ways like Jackson's. It concerns not assertability [Mellor, 1993, p. 234 fn. 6], but acceptability: to accept a conditional is to be disposed to infer its consequent from its antecedent; my degree of acceptance of 'If *A*, *B*' is the degree of belief I am disposed to have in *B* if I fully believe *A*. This is close enough to the Thesis: any peripheral differences are not my present concern (see above, pp. 159–160). To accept a conditional is not to have a belief, but to

have an inferential disposition. But he argues, this does not deprive conditionals of truth conditions. What Lewis showed [1976] is that there is no proposition, no object of belief, such that your degree of belief in its truth systematically matches your conditional degree of belief in *B* given *A*. Once we have decided that to accept a conditional is not to believe something, Mellor claims, Lewis's result becomes irrelevant to the question whether conditionals have truth conditions.

Mellor reminds us that beliefs are not the only mental states that have truth conditions. Desires, fears, and other propositional attitudes have them too. So why shouldn't the dispositional states which constitute acceptance of conditionals also have them? 'All Lewis shows is that an Adams 'If *P*, *Q*' cannot express a *belief* in . . . [its truth-conditional] content. So nothing stops the [dispositional] theory crediting all. . . conditionals with. . . truth conditions' [Mellor, 1993, p. 238].

Mellor thinks there are two kinds of conditionals, and has been convinced by Dudman that the traditional line was misplaced ([Dudman, 1988]; and see above, p. 129). Backward-looking conditionals like 'If Oswald didn't do it, someone else did', have truth-functional truth conditions. Forward-looking conditionals, like 'If Oswald doesn't do it, someone else will', behave like those traditionally called 'subjunctive'. These, Mellor suggests, have Stalnaker–Lewis-style truth conditions.

Take the truth-functional case. Belief that the truth condition is satisfied is not enough to accept the conditional, for well-known reasons. But if you do accept 'If *A*, *B*', the truth of $A \supset B$ is what ensures that you won't end up with a false belief in *B*, should you learn *A*.

That is so, but inferring the consequent from the antecedent is not the only thing we do with conditionals we accept. They have other roles in practical and theoretical reasoning. And accepting truth-functionally true ones could get you into all sorts of trouble. Being in a fragile state of mind, I *accept* that if the Queen was at home this last hour, she has been worrying about where I am. So I had better try and phone. I am liable to be arrested for making nuisance calls, though the conditional I accept is true, for she is not at home. Had the Warren Commission accepted 'If Oswald didn't kill Kennedy, MI5 did', Anglo-American relations would have sunk to an all-time low, even if their conditional was true (Oswald did it). We should try not to accept conditionals which are truth-functionally false, everyone agrees; but accepting a conditional can be a pretty bad inferential disposition to have, even if its material counterpart is true.

The converse problem arises for the conditionals to which Mellor ascribes strong truth conditions: I may be disposed to have a high degree of belief in *B* on learning *A*, yet be fairly sure that 'If *A*, *B*' is not true. To repeat the boring old example, I am disposed to a high degree of belief in 'the coin will land heads at least once' should I acquire the belief that you are going to toss it ten times. So I accept the corresponding conditional. I also happen to think it's unlikely that you will toss it ten times. On Lewis's and Stalnaker's truth conditions, it is true only if you do toss it ten times and get at least one head — and this is unlikely, for it is unlikely that you will toss it. On Lewis's truth conditions, it is otherwise false — either you toss

it and get no heads, or you don't toss it, but there are *some* close worlds in which you toss it and get no heads. On Stalnaker's, it is false if you toss and get no heads, indeterminate if you don't toss it. For both, the probability of its *truth* is low. (I use the coin example because its structure is transparent; but many everyday examples produce the same result.) Mellor misses this discrepancy because, although he allows that there are degrees of acceptance and of belief, he does not consider their application in his discussion of possible-world truth conditions. The discrepancy between his acceptance condition and the truth condition does not show up in the case of full belief. If I'm certain that all relevant *A*-worlds are *B* worlds, I will fully believe *B* if I learn *A*, and vice versa. The discrepancy shows up if we replace 'all' with 'almost all'.

It is not clear what role the truth conditions play when they fit the acceptance condition badly (as, of course, they must, given Lewis's result). Mellor reminded us that desires, hopes, fears also have truth conditions. So they do: to desire that *p* is to desire that it is true that *p*, i.e., that *p*'s truth condition obtains. That is to say, these states have propositional content. But a conditional does not have a propositional content, according to Mellor, rather, 'it has not one content, but two, namely *P* and *Q*' [Mellor, 1993, p. 238]. To accept a conditional is not to accept that it is true, that its truth condition obtains. And yet it has a truth condition — one which I may believe obtains while not accepting the conditional, or one which I may disbelieve while accepting the conditional. Now 'has a propositional content' and 'has a truth condition' are somewhat technical terms, but we use them interchangeably. It has been hard enough to get our minds round the idea that conditionals have neither. It is harder still, I think, to accept that they have one but not the other.

9.3

Conditionals may have truth conditions which are radically context dependent. Van Fraassen's complaint against Lewis's proof was the assumption that a conditional will express the same proposition in different belief states: 'the logical disaster was precipitated not by Stalnaker's Thesis [the equation of the probability of a proposition with a conditional probability], but by [Stalnaker's] Thesis coupled with Lewis's metaphysical realism' [van Fraassen, 1976, p. 275]. Lewis had said 'presumably our indicative conditional has a fixed interpretation, the same for speakers with different beliefs, and for one speaker before and after a change in his beliefs. Else how are disagreements about a conditional possible, or changes of mind?' [Lewis, 1976, p. 138]. Stalnaker showed that even in a single belief distribution, the Equation cannot hold for all conditionals (see Section 6.4 above); but, for a simple conditional with no embedded conditionals, we can, perhaps, always find some proposition for it to express in a given belief distribution. Stalnaker [1975] argued for a context-dependent interpretation of his truth conditions for indicative conditionals. Later he said that 'to play their methodological role, [indicative] conditionals must be too closely tied to the agents who utter them for

those conditionals to express propositions which could be separated from the contexts in which they are accepted' [Stalnaker, 1984, p. 111].

Stalnaker [1975] addresses a problem which is the mirror image of that addressed by Grice and Jackson. In Section 6.5, p. 169, I mentioned two prima facie desirable properties of indicative conditionals: (i) minimal certainty that $A \vee B$ (ruling out just $\neg A \& \neg B$) is enough for certainty that if $\neg A, B$; and (ii) it is not necessarily irrational to disbelieve A yet disbelieve that if A, B. I showed that no proposition can satisfy both. The truth-functional conditional satisfies the first and not the second. Grice and Jackson provide a surrogate for the second: if you disbelieve A, then, although $\neg A$ entails 'If A, B', the latter may still be *unassertable*. Stalnaker's conditional satisfies the second but not the first. He argues for a surrogate of the first. Although the inference

> Either the butler or the gardener did it. Therefore, if the butler didn't do it, the gardener did

is invalid on his semantics, nevertheless, whenever the first is *assertable*, so is the second. Like Grice, Stalnaker appeals to the pragmatics of communication.

Stalnaker's formal semantics uses a 'selection function', f, which selects, for any proposition A and any world w, a world, w', the nearest (most similar) world to w at which A is true. 'If A, B' is true at w iff B is true at $f(A, w)$, i.e. at w', the world most similar to w at which A is true. 'If A, B' is true simpliciter iff B is true at the nearest A-world to the actual world. (However, we do not know which world is the actual world. To be sure that if A, B, we need to be sure that whichever world w is a candidate for actuality, B is true at the nearest A-world to w.) If A is true, the nearest A-world to the actual world is the actual world itself, so in this case 'If A, B' is true iff B is also true. The selection function does substantive work only when A is false.

In the case of indicative conditionals the selection function is subject to a pragmatic constraint, set in the framework of the dynamics of conversation. At any stage in a conversation, many things are taken for granted by speaker and hearer, i.e. many possibilities are taken as already ruled out. The remaining possibilities are live. Stalnaker calls the set of worlds which are not ruled out — the live possibilities — the context set. For indicative conditionals, antecedents are typically live possibilities, and we focus on that case. The pragmatic constraint for indicative conditionals says that if the antecedent A is compatible with the context set (i.e. true at some worlds in the context set) then for any world w in the context set, the nearest A-world to w — i.e. the world picked out by the selection function — is also a member of the context set. Roughly, if A is a live possibility (i.e. not already ruled out), then for any world w which is a live possibility, the nearest A-world to w is also a live possibility.

The proposition expressed by 'If A, B' is the set of worlds w such that the nearest A-world to w is a B-world. The ordering of worlds, by the pragmatic constraint, depends on the conversational setting. As different possibilities are live in different

conversational settings, a different proposition may be expressed by 'If A, B' in different conversational settings.

Consider the one-person case: I am talking to myself, i.e. thinking — deliberating about whether if A, B. The context set is the set of worlds compatible with what I take for granted, i.e. the set of worlds not ruled out, i.e. the set of worlds which are epistemically possible for me. Let A be epistemically possible for me. Then the pragmatic constraint requires that for any world in the context set, the nearest A-world to it is also in the context set. Provided you and I have different bodies of information, the proposition I am considering when I consider whether if A, B may well differ from the proposition you would express in the same words: the constraints on nearness differ; worlds which are near for me may not be near for you.

Stalnaker assumes that a disjunction, '$A \vee B$' is assertable only if $A\&\neg B$ and $\neg A\&B$ are live possibilities, but $\neg A\&\neg B$ has been ruled out. Hence, if the disjunction is assertable, the context set contains some $\neg A\&B$-worlds and no $\neg A\&\neg B$-worlds. So all the $\neg A$-worlds in the context set are B-worlds. So whichever world in the context set is actual, the nearest $\neg A$-world to it is a B-world. Hence, in a context in which '$A \vee B$' is assertable, so is 'If $\neg A, B$'.

Thus Stalnaker avoids the argument against non-truth-functional truth conditions given in 6.5. The argument may be spelled out as follows. There are six incompatible logically possible combinations of truth values for A, B and $\neg A \rightarrow B$:

	A	$\neg A$	B	$A \vee B$	$\neg A > B$
1.	T	F	T	T	T
2.	T	F	T	T	F
3.	T	F	F	T	T
4.	T	F	F	T	F
5.	F	T	T	T	T
6.	F	T	F	F	F

We start off with no firm beliefs about which obtains. Now we eliminate just $\neg A\&\neg B$, i.e. establish A or B. That leaves five remaining possibilities, including two in which '$\neg A > B$' is false. So we can't be certain that $\neg A > B$ (whereas, intuitively, one can be certain of the conditional in these circumstances). Stalnaker replies: we can't, indeed, be certain that the proposition we were wondering about earlier is true. But we are now in a new context: $\neg A\&\neg B$-worlds have been ruled out (but $\neg A\&B$-worlds remain). We now express a different proposition by '$\neg A > B$', with different truth conditions, governed by a new nearness relation. As all our live $\neg A$-worlds are B-worlds (none are $\neg B$-worlds), we know that the new proposition is true.

Now this hypersensitivity of the proposition expressed by 'If A, B' to what is taken for granted by speaker and hearer, or to the epistemic state of the thinker, is not very plausible. One usually distinguishes sharply between the content of what is said and the different epistemic attitudes one may take to that same content.

Someone conjectures that if Ann isn't home, Bob is. We are entirely agnostic about this. Then we discover that at least one of them is at home (nothing stronger). We now accept the conditional. It seems more natural to say that we now have a different attitude to the same conditional thought, that B on the supposition that $\neg A$. It does not seem that the content of our conditional thought has changed. And if there are conditional propositions, it seems more natural to say that we now take to be true what we were previously wondering about. There does not seem to be any independent motivation for thinking the content of the proposition has changed.

Also, Stalnaker's argument is restricted to the special case where we take the $\neg A\&\neg B$-possibilities to be ruled out. Consider a case when, starting out agnostic, we become close to certain, but not quite certain, that A or B — say we become about 95% certain that A or B, and are about 50% certain that A. According to the Thesis, we are entitled to be quite close to certain that if $\neg A$, B — 90% certain in fact. (If $b(A \text{ or } B) = 95\%$ and $b(A) = 50\%$, then $b(\neg A\&B) = 45\%$. $b(\neg A\&\neg B) = 5\%$. So, on the assumption that $\neg A$, it's 45:5, or 9:1, that B.) In this case, no additional possibilities have been ruled out. There are $\neg A\&\neg B$-worlds as well as $\neg A\&B$-worlds which are permissible candidates for being nearest. Stalnaker has not told us why we should think it likely, in this case, that the nearest $\neg A$-world is a B-world.

Uncertain conditional judgements create difficulties for all propositional theories. As we have seen, it is easy to construct probabilistic counterexamples to the truth-functional theory; and it is easy to do so for the variant of Stalnaker's theory according to which 'If A, B' is true iff B is true at *all* nearest A-worlds (as Lewis [1973]) holds for counterfactuals). (It is very close to certain that if you toss the coin ten times, you will get at least one head; but it is certainly false that the consequent is true at all nearest antecedent-worlds.) It is rather harder for Stalnaker's theory, because nearness is so volatile, and also because it is not fully specified. Here is a putative counterexample. (I owe this example to a student, James Studd, who used it for a slightly different purpose.)

We have no idea how much fuel, if any, there is in the car (the gauge isn't working). Ann is going to drive it at constant speed, using fuel at a uniform rate, along a road which is 100 miles long. The capacity of the tank is just enough to do 100 miles: if the tank is full she will go 100 miles then stop. If the tank is $x\%$ full, she will go x miles then stop. We give equal credence to the propositions 'She'll stop in the first mile', 'She'll stop in the second mile' and so on.

Now consider the conditionals

(1.) If she stops before half way, she will stop in the 1st mile.

. . .

(50.) If she stops before half way, she will stop in the 50th mile.

According to the Thesis, these are all equally likely — each is 2% likely. This seems reasonable.

Write Stalnaker's truth condition thus:

> '$A > B$' is true iff either $A\&B$, or $\neg A$ and the nearest A-world is a B-world.

The following assumption is very plausible: consider a world w in which Ann goes more than half way. The most similar world to w in which she does not go more than half way is one in which she stops in the 50th mile. After all, it is spatially and temporally more similar, more similar in terms of the amount of fuel in the tank, more similar in its likely causes and consequences, etc., than a world in which she stops earlier.

Let us evaluate (1) and (50) using Stalnaker's truth condition. There are two ways in which (1) can be true: (a) she stops in the first mile (1% likely); (b) she doesn't stop before half way and in the nearest world in which she does stop before half way she stops in the first mile. By our assumption, (b) is certainly false. So (1) has a probability of 1%.

There are two ways in which (50) can be true: (a) she stops in the 50th mile (1% likely); (b) she doesn't stop before half way and in the nearest world in which she does stop before half way she stops in the 50th mile. By our assumption, (b) is true iff she doesn't stop before half way, and so is 50% likely. So (50) gets a total probability of 51%.

So, given the plausible assumption about nearness, Stalnaker's theory gives implausible answers to (1) and (50). The example also shows how unnatural the thought-experiment is, as an assessment of how likely it is that if A, B, using Stalnaker's truth conditions for 'If A, B'.

In Stalnaker's defence, perhaps the assumption should be rejected: in the context, there is nothing to choose between worlds such as (1) and (50), regarding how close they are to a world w in which Ann goes more than half way. Then the selection function does not have a definite value for this argument, and as the conditionals are not true for all permissible values, they are indeterminate (see [Stalnaker, 1981]). So, in a sense, they are, but the verdict 'indeterminate' is not very informative. And we do not have a worked-out theory of uncertainty about indeterminate propositions. Some ideas for such a theory are explored in the next section, §9.4.

9.4

Van Fraassen [1976, pp. 279–282] had an idea for adapting Stalnaker's semantics so that your degree of belief in a Stalnaker conditional $A > B$ equals your $b(B \text{ given } A)$. Closely related ideas are found in [McGee, 1989; Jeffrey, 1991] and [Stalnaker and Jeffrey, 1994]. The proposition $A > B$ will not be independent of your belief state, but it will yield a theory of what you should believe about compounds of conditionals. This has been the focus of recent work.

As we saw above, Stalnaker's formal semantics is equipped with a selection function, f, which selects, for any world w and any proposition A, the 'nearest'

world w' to w in which A is true. Let w be the actual world. If A is actually true f selects the actual world. If A is actually false, f selects the 'nearest' A world. $A > B$ is true iff B is true at the world f selects for A. Now suppose you think it's 80% likely that B given A. For expository purposes only, let me express this: you think 80% of the A-worlds are B-worlds. In the $A\&B$-worlds, $A > B$ is true. In the $A\&\neg B$-worlds, $A > B$ is false. If the actual world is a $\neg A$-world, is it one for which f selects an $A\&B$-world, or one for which f selects an $A\&\neg B$-world? Well, you don't know; and there may be no determinate answer to the question, for there may be nothing to choose between different A-worlds. Stalnaker never did believe that there were hard facts about which worlds were 'nearest', or how actual selection functions work: this was his way of 'projecting epistemic strategies onto the world' (see p. 163 above). The best projection, Van Fraassen suggested, would be this. If A is false, *let the selection function select an A-world at random*. Then how likely is it (for you) that it selects a B-world? 80%, because you think 80% of the A-worlds are B-worlds. So $A > B$ has an 80% probability of being true if an A-world obtains, and an 80% probability of being true if a $\neg A$-world obtains; so an 80% probability of being true. $b(A > B) = b(B|A)$.[66]

Suppose 90% of the red balls have black spots. How likely is it that, if you pick a red ball (R), it will have a black spot (B)? Your $b(B|R) = 0.9$. $R > B$ is true if $R\&B$, false if $R\&\neg B$; if $\neg R$ there is a 90% chance that an $R\&B$-world is 'selected' and $(R > B)$ is true. So $b(R > B) = 0.9$. You think it's very likely that if they are at home (H), the lights will be on (L). Suppose they are not at home. Then the selection function is very likely to select an $H\&L$-world rather than an $H\&\neg L$-world. So $b(H > L)$ is high.

This is to give up genuine truth values for the conditional when its antecedent is false. The $\neg A$-worlds don't really divide into those in which $A > B$ is true and those in which $A > B$ is false. $A > B$ is indeterminate in all the $\neg A$-worlds (when your $b(B|A)$ is neither 1 nor 0). This would, I think, block Stalnaker's version of the bombshell (see Section 6.4). His proof did assume that the $\neg A$-worlds divide into the $\neg A\&(A > B)$-worlds, and the $\neg A\&\neg(A > B)$-worlds.

Jeffrey [1991] got the same effect by giving 'If A, B' an intermediate 'semantic value' equal to your $b(B|A)$ when A is false. The conditional is, as it were, 80% true if $\neg A$, when your $b(B|A)$ is 0.8. If we write '1' for 'T' and '0' for 'F', we get a 'truth table' for the conditional that looks like this

b	A	B	If A, B
0.4	1	1	1
0.1	1	0	0
0.5	0		0.8

Extending the notion of degree of belief to the case where the object of belief is a three-valued entity, he takes the weighted average of its semantic value. For

[66] Van Fraassen proved far from trivial results showing how to apply this idea to an infinite set of worlds.

degrees of belief in $A\&B$, $A\&\neg B$ and $\neg A$ as above, we get $b(\text{if } A, B) = (0.4 \times 1) + (0.1 \times 0) + (0.5 \times 0.8) = 0.8 = b(B|A)$.

Jeffrey [1991] devised ways of assigning degrees of belief to compounds of conditionals from this basis. Stalnaker and Jeffrey [1994] show that this construction is equivalent to Van Fraassen's. *Within a given belief distribution*, a conditional degree of belief can be equated with a degree of belief (in this extended sense) in a three-valued entity, and degrees of belief in compounds of conditionals can be assigned.

It is, admittedly, a rather weird three-valued entity. The 1 and 0 are truth values. The 0.8 is a degree of belief. This is an odd mixture of ingredients in a weighted average. Replying to Jeffrey [Edgington, 1991, pp. 203–205], I thought we could get the same effect by taking the third value to be the objective probability of B given A (where this exists); and that our epistemic estimation of this objective three-value identity would still be our $b(B|A)$. But I was wrong. The bombshell extends: there is no three-valued entity such that, in *all* belief distributions, your epistemic estimate of its value is your $b(B|A)$.

McGee [1989] tackled directly the question of assigning probabilities to e.g. conjunctions and negations of conditionals whose values are conditional probabilities. Part of his methodology was to investigate what would be fair betting odds on conjunctions of conditionals.

If we can find a general way of assigning probabilities to conjunctions and negations of conditionals, we have a means of assigning truth values to them (and, more obviously, vice versa). Start with a set of conditionals. Form from it a set of 'state-descriptions' — conjunctions which contain, for every conditional, either it or its negation. These form a partition. They are surrogate possible worlds. Suppose we can assign probabilities to these, which sum to 1. The probability of a conditional should be the sum of the probabilities of the state-descriptions in which it is 'true', viz. unnegated (see [Adams, 1975, pp. 32–33]). McGee ends by showing that on his construction conditionals which satisfy the Thesis can be construed as Stalnaker-like conditionals with a random selection function, like Van Fraassen's.

I say 'Stalnaker-like' because McGee makes one modification. An important part of McGee's construction is the equivalence of 'If $A\&B$, then C' and 'If A, then if B then C'. This is invalid on Stalnaker's semantics (see above, p. 172). But only a small change is needed to modify the semantics in this respect. This is one difference between Jeffrey's and McGee's constructions. Another, connected difference is that Jeffrey applies his methods to conditionals in antecedents of conditionals, and McGee does not: McGee's antecedents are always 'factual' sentences. Apart from these differences, they get the same results, by different methods.

Unfortunately, their results about compounds of conditionals are not altogether pleasing. Lance [1991] has a plausible counterexample to their common account of conjunctions of conditionals. I raised some further difficulties [Edgington, 1991, pp. 200–202]. Stalnaker and Jeffrey [Stalnaker and Jeffrey, 1994] and McGee (private communication) concede that their theories have counterintuitive consequences. Intuition, though, is not very robust on this subject. We seem to know

enough about compounds of conditionals to reject certain claims (see Section 7.1 above). Others remain controversial, and some we don't know how to understand. This is an area in which there is more work going on, and there may be more to be learned.

It is hard to decide whether there is more to be learned. Conditionals, these theorists concede, are not ordinary propositions, so there is no a priori reason why there should be general routines for decoding compounds of them. On the other hand, they are not that different from ordinary propositions; and they stand in definite relations to ordinary propositions ($A\&B$ entails 'if A, B', which entails $A \supset B$, on Adams' account). In Section 7.1 I agreed with Gibbard that, with the help of a context, we adopt ad hoc strategies for finding suitable interpretations. Sometimes, without a suitable context, we fail to understand them. These general theories give the impression of going beyond the data, and going beyond our practical needs. But a general theory which did not clash with those intuitions we do have would be an achievement, one for which it is likely we would find a use.

10 COUNTERFACTUALS (BY ANY OTHER NAME)

10.1

We return to 'counterfactual' or 'subjunctive' conditionals — those expressed in English with a 'would' in the consequent — and their relation to indicatives. (As mentioned in §1.2, neither of the standard names is entirely appropriate: 'counterfactuals' need not be literally counterfactual; and 'subjunctive' is controversial — Stalnaker [2005] speaks of 'the combination of tense, aspect and mood that we have gotten into the habit of calling "subjunctive"'.) Consider again these three forms:

(1a) If she caught the 10 o'clock train, she arrived at noon;

(1b) If she catches the 10 o'clock train, she will arrive at noon;

(1c) If she had caught the 10 o'clock train, she would have arrived at noon.

What is the relationship between them? Intuitively, very close: you would not take yourself to have expressed a radically new kind of thought, having passed from one to another in a suitable context.

Those who defend the text-book truth-functional theory of the indicative conditional must see the so-called subjunctive conditional as radically different, and indeed logically stronger than the indicative conditional — as Jackson and Lewis do. The difference, it is alleged, is shown by the OK cases:

(2a) If Oswald didn't kill Kennedy, someone else did;

(2c) If Oswald hadn't killed Kennedy, someone else would have.

You may accept the former and reject the latter. But the thought that the subjunctive is logically stronger is undermined when we switch the consequents:

(3a) If Oswald didn't kill Kennedy, no one else did;

(3c) If Oswald hadn't killed Kennedy no one else would have.

Suppose you accept (2a) and reject (2c); then you reject (3a) and accept (3c): one may accept the subjunctive and reject the corresponding indicative. So the former cannot be logically stronger than the latter; rather, the two forms would appear to be independent in strength.[67]

Those who take the ordinary indicative conditional also to be logically stronger than the truth-functional conditional cite in their favour the advantage of being able to provide a broadly unified theory of conditionals, within which one hopes to explain the difference made by the change in grammatical form. This perspective goes back a long way. Ayers [1965] opens a paper in *Mind* with the remark that there is no special problem of subjunctive, or counterfactual conditionals, there is just a problem of conditionals. The same thought informs Strawson's writings both in *Introduction to Logical Theory* [1952] and in a later paper [Strawson, 1986], quoted above (p. 11). And this thought lies behind Stalnaker's approach to the subject.

As mentioned above (pp. 156–157), Adams [1975, chapter 4; 1993] held that the Thesis could be extended to subjunctive conditionals: while differences must be accounted for, conditional probability is the key to the assessment of subjunctive conditionals also. This has not been a popular view. Many philosophers follow Adams on indicatives but take a Lewis-Stalnaker line on subjunctives, for instance Gibbard [1981], Appiah [1985], and Bennett [2003]. When Bennett turns to subjunctives an entirely new battery of machinery arrives on the scene, and he says 'they are not in any deep way like indicatives' (ibid. p. 256). Here he echoes Gibbard, who wrote that 'the apparent similarity between these two "if" constructions hides a profound semantic difference' (ibid. p. 211). Indeed there is a profound semantic difference on this view: subjunctive conditionals express propositions with truth conditions, indicative conditionals do not. Can the difference be so great between (1b) and (1c) above?

To be fair to Gibbard, he treats the 'will' conditional as ambiguous: it can be read as indicative or subjunctive [Gibbard, 1981, p. 228]. Thus he subscribes to a milder version of the 'relocation thesis' (Bennett's term) than Dudman's, who holds that the 'wills' and 'woulds', like (1b) and (1c), are one kind of conditional; those that are neither 'wills' nor 'woulds', Dudman calls 'hypotheticals', and treats quite differently. However, intuitively, (1a) above does not seem to express a radically different kind of thought from (1c), either.

[67] This is yet another argument against the truth-functional interpretation of the indicative conditional: the truth-functional conditional is weaker than, and entailed by, the corresponding subjunctive. But the indicative conditional may be rejected while the corresponding subjunctive is accepted. Therefore, the truth-functional conditional is not the indicative conditional.

We shall consider the relocation thesis later. I turn to the question whether, and if so how, the Thesis may be extended to subjunctives.

10.2

According to the Thesis, an indicative conditional statement is not a categorical statement of a proposition, true or false as the case may be; it is rather a statement of the consequent under the supposition of the antecedent. A conditional belief is not a categorical belief that something is the case; it is a belief in the consequent in the context of a supposition of the antecedent. Note that the easy transition between 'suppose' and 'if' is as natural for subjunctives as it is for indicatives: suppose they had been at home; then the lights would have been on. Suppose Kennedy had not been assassinated; then there would have been no Vietnam war.

The strongest evidence for the Thesis comes from considering uncertain conditional judgements. Uncertainty is as prominent a feature of subjunctives as it is of indicatives: the doctor is close to certain, but not quite certain, that the patient will be cured if he has the operation. The operation is declined, and on the same grounds the doctor is close to certain that he would have been cured if he had had the operation. If conditional probability gives the structure of the former judgement, it would seem that it is also the key to the latter. But while both backward-looking and forward-looking indicative conditional judgements reflect our degree of confidence in the consequent given the antecedent, for subjunctive conditional judgements, the relevant conditional probability does not (normally) reflect your current degree of confidence in C given A, but (most typically) your view about how likely it *was* that C would have happened, given that A had.[68]

10.3

Let us take a step back and see how these past-tense probability judgements arise, first in their unconditional form. Consider our judgements about the future. Often the future cannot be known, but sometimes there is something more accessible to guide our uncertain judgements, for there are objective chances of future events, which can be estimated more or less accurately. We consult experts. Some people do experiments to estimate the chances. Whether or not there are objective chances at the most fundamental level of description of the world, there are stable, discoverable — within limits — features of the world which generate the sorts of patterns we would expect if there are chances; and at usable levels of description, they are ubiquitous.

When I say our uncertain judgements about the future are — or better, should be — guided by the chances, I am appealing in part to what David Lewis [1980]

[68]The parenthetical qualifications are there because (a) sometimes, about the future, it might be a matter of indifference whether one says 'If it rains…' or 'If it were to rain…'; and (b) not every context shift away from your actual epistemic state is a temporal one: 'Suppose Euclidean geometry had been true of the real world…'.

named the Principal Principle: beliefs about chances rationally constrain credence or degree of belief about future outcomes. One cannot rationally be convinced that the coin is fair (i.e. is such that the objective chance of its landing heads is 0.5) yet be close to certain that it will land heads on this toss. But what I am claiming goes beyond the Principal Principle, and in a sense explains it: in matters of objective chance about the future, the best degree of confidence to have which is attainable by reliable and rational means, is one that matches the objective chance, and this is something our beliefs sensibly aim at. Why else would it be irrational to be sure that the coin is fair yet close to certain that it will land heads on this toss?

With our uncertain judgements about the past, chance has no such authoritative normative role for belief. For the chances have already played themselves out — settled down to 1 or 0. It *may* be the best you can do to base your judgement about something in the past on your knowledge of what the chance *was* that it *would* happen, e.g. that the coin was fair when it was tossed. But it may not be; for we can have information downstream from the outcome making it certain, or close to certain, that it landed heads. We can judge that it was probable that such and such *would* happen, but it didn't, or it is now unlikely that it did.

Chances change with time: when was it probable that such-and-such would happen? Context may make this clear enough (as with the past tense in general), or the time can be specified. It was probable at the beginning of the game that Spurs would win; throughout the second half it became less and less probable, but lo and behold, in the last minute they equalised and in extra time they won.

What does it mean to judge that it *was* probable that something *would* happen? It doesn't mean that you were close to certain that it would. You might have been, or you might not have been. You can say 'It was very probable that such-and-such would happen, though I didn't realise it at the time'. And indeed, the event in question might concern a time before you were born — it could even concern a time before anyone was born, say if it was about the survival chances of dinosaurs. It can mean that there was, at the time in question, a high objective chance that such and such would happen. Hence, you are endorsing the corresponding hypothetical degree of confidence that would have been the right one at the earlier time. And it is hard to see what else it could mean.

For 'It was probable that such-and-such would happen', it's not necessary (as I said) that you had an earlier high degree of confidence; nor is it sufficient. Suppose that people have spun lies to me that I stand a good chance of making a fortune by investing in this concern. I hand over my money, and it disappears without trace. I realise I have been conned. Again, I don't felicitously say 'It was very probable that I would make a fortune', but 'I thought it was very probable that I would make a fortune'.[69] In other words, there is a drive towards objectivity. And when we are talking about an earlier time, we form judgements about how the chances were at

[69] I think the same is true of the epistemic 'might'. 'It might land heads', I say. I later discover that it was a double-tailed coin. I don't felicitously say 'It might have landed heads' but rather 'I thought it might have landed heads'.

that time. The best sense we can make of the locution 'It *was* likely that such-and-such *would* happen' is as a judgement about how the objective chances were at some past time.

All these sorts of judgements can occur within the context of a supposition. We make suppositions when deliberating about what to do, and we make suppositions when deliberating about what is the case. You can be more or less certain of these judgements in the context of a supposition: that Jane will accept on the supposition that she is offered the job; that Smith took the money on the supposition that Jones didn't; that the dog would have bitten me on the supposition that I had approached, and so on. And suppositions (according to the Thesis) are more succinctly formulated with the word 'if'.

Our backward-looking and forward-looking indicative conditional judgements reflect our degree of confidence in the consequent on the supposition of the antecedent. With the backward-looking ones, we all have our own idiosyncratic combinations of knowledge and ignorance, and, as Gibbard [1981] made famous, people can faultlessly come to opposite opinions: there is nothing objective to aim at. For instance, we both start off knowing that X or Y or Z did it. I discover that it wasn't Y. You discover that it wasn't Z. I accept 'If not X, Z' and reject 'If not X, Y'. You do just the opposite. All the relevant facts are available, but neither of us has them all, and we would know if we pooled our information that X did it. The present objective chance that X didn't do it is zero, so there is no such thing as the present objective chance of Y supposing that not X (just as there is no such thing as an objective chance that you will pick a spotted ball if you pick a red ball, when there are no red balls in the bag). For many forward-looking indicatives, by contrast, it is not yet determined whether the antecedent is true (or at least we treat it as not yet determined, not yet knowable, whether the antecedent is true), and there may be such a thing as the objective conditional chance of C given A for our judgements to aim at. And after the event, should it turn out that not A, we can be right or wrong when we say that, very likely, C would have happened if A had — very likely, you would have picked a spotted ball if you had picked a red ball (assuming now that there were some red balls in the bag). Thus, it is hoped, we get a fundamentally unified account of conditional judgements, which also explains interesting differences between the different kinds. Of course we may be certain that if A, C, but typically we are not, and conditional probability — the probability of C on the supposition that A — is the key to how close to certain we are of a conditional of any kind. But, as already remarked, in the case of subjunctive conditionals, this conditional probability does not typically represent your current degree of belief in C given A.

Note that we secure the required independence in the Oswald-Kennedy cases: it is obviously consistent to have a high degree of belief that someone else did it on the supposition that Oswald didn't (i.e. a low degree of belief that no one else did it on the supposition that Oswald didn't), but to judge that the probability was low, back then in 1963, that someone else would have killed Kennedy, supposing that Oswald hadn't (i.e. the probability was high, back then in 1963, that no one

else would have, supposing that Oswald hadn't).

Why do these judgements matter — judgements about what would have happened if…? Here is one reason which I am inclined to think is the principal one: they play an indispensable role in empirical reasoning about what is the case. In abduction, or inference to the best explanation, we look for hypotheses such that what we do observe is what we *would* expect to observe if it were the case that H, and would not expect to observe, were it not the case that H. 'It's not a problem with the liver' says the doctor, 'for the blood test was normal; and if it had been a problem with the liver, it would have been such-and-such'. 'They're not at home; for the lights are off; and if they had been at home, the lights would have been on'; 'I think the patient took arsenic; for he has such-and-such symptoms; and these are the symptoms he would have if he had taken arsenic.' 'I think the prisoner jumped from that window; for the flowers below are squashed; and they would have been squashed if he had jumped from there'. These are not intended as deductively valid arguments. They can be defeated if reasonable alternative hypotheses make it just as likely, or unlikely, that we would have observed what we do observe. For instance it may be pointed out that the flowers would also be squashed if there had been a game of football, or a dog fight. Or it could be pointed out that they always leave the lights on when they go out at night, so there must be some other explanation of the lights being off — they have gone to bed early, or there was a power cut. Nevertheless, they are part and parcel of the most basic kind of empirical reasoning. Nor are the conditionals involved typically certain rather than probable. But to the extent that we can find a hypothesis H such that the chance was high that we would observe E (as we actually do), on the supposition that H is true, and the chance was low that E given $\neg H$, and H is not initially too unlikely, we have a good argument for H.

10.4

Now, a conditional probability — the probability of C on the supposition that A — is not a measure of the probability of the truth of a proposition. There is no proposition X such that, necessarily, the probability that X is true is the conditional probability of C given A. If subjunctives are to be understood in terms of conditional probabilities, they are not to be understood in terms of truth conditions. For if they are to be understood in terms of truth conditions, you should believe a subjunctive to the extent that you think it is true — that its truth conditions are satisfied.

Bennett [2003, pp. 254–6] claims that none of the arguments against truth conditions for indicative conditionals work for subjunctive conditionals because they all have at least one false premiss. For example, an argument I have given (§6.4, see p. 169) includes the premiss that if you are certain that A or B without being certain that A, you must be certain that if $\neg A, B$. But this is false for subjunctives. To take the most famous example, I can be certain that either Oswald killed Kennedy or someone other than Oswald killed Kennedy (while less than certain

that Oswald did it), without being certain that if Oswald hadn't killed Kennedy someone else would have. But Bennett misses the point that the conditional probabilities relevant to the assessment of subjunctive conditionals do not (typically) represent your present actual distribution of belief, but those of a hypothetical belief state in a different context, normally that of an earlier time, concerning (e.g.) whether someone else will kill Kennedy if Oswald doesn't. And it is on that earlier belief state that one would run the arguments. One could say: your attitude to the subjunctive endorses the hypothetical matching attitude to the earlier indicative; and we could run the standard arguments to show that the attitude to the earlier indicative is not an attitude to a proposition.[70]

The standard arguments were aimed at indicative conditionals. They apply the general structural fact that a conditional probability does not measure the probability of the truth of a proposition, to the case where the conditional represents your actual present state of conditional belief.

Here is another way of looking at it: the fact that a conditional probability is not the probability of the truth of a proposition is in a sense the same structural fact as the fact that quantifiers like 'most', 'almost all' in 'Most *A*s are *B*', 'Almost all *A*s are *B*', or '90% of *A*s are *B*', unlike the standard treatment of the quantifiers 'all' and 'some' in 'All *A*s are *B*' and 'Some *A*s are *B*', are essentially binary, restricted quantifiers, in that they cannot be reduced to unary, unrestricted quantifiers 'Most things [in the domain] are…'. For probability statements can be modelled by statements about proportions. Let me divide logical space into a finite number of (in my judgement) equiprobable bits, adequate for the problem at hand, i.e. every proposition I am concerned with is true throughout, or false throughout, any bit. For the sake of familiarity I shall call the bits 'worlds'; though they are not ultimate not-further-subdividable possibilities, they are divided finely enough for the project at hand. A proposition *B* is probable iff it is true in most of the worlds; it is almost certain iff it is true in almost all of the worlds; it is 90% probable iff it is true in 90% of the worlds. A proposition *B* is conditionally probable, on the supposition that *A*, iff most *A*-worlds are *B*-worlds; almost certain if almost all *A*-worlds are *B*-worlds; 90% probable if 90% of the *A*-worlds are *B*-worlds. If these were equivalent to statements about the probability of some proposition *X*, they would be equivalent to something of the form: in most worlds, *X* is true; most worlds are *X*-worlds; almost all worlds are *X*-worlds; 90% of worlds are *X*-worlds, etc.; and we would have expressed the 'most' in 'Most *A*s are *B*' as a unary quantifier, which cannot be done.[71]

[70] Actually Bennett's claim is not true of Lewis's first proof [Lewis, 1976], which is just that there is no proposition such that necessarily, the probability of its truth is the conditional probability of *C* given *A*.

[71] In the general theory of quantifiers the first predicate is sometimes called the 'restrictor'; that describes what the antecedent of a conditional does: it restricts the claim that *C* to a context in which the antecedent, *A*, is satisfied.

10.5

I say 'If you touch that wire you will get a shock'. You don't touch it. I use my circuit-testing instrument to show you: 'You see, if you had touched it you would have got a shock'. Or, if the result is different: 'Funny, the power must be off. I was wrong. You wouldn't have got a shock if you had touched it'. A dog almost always, but not quite always, attacks and bites when strangers approach. I'm told 'It's very likely that you will be bitten if you approach'. I don't approach. Trusting my informant, I say 'It's very likely that I would have been bitten if I had approached'. Fred asks his doctor if he will be cured if he has the operation. The doctor says 'We can't be sure, but I'm pretty sure — about 90% sure that you will be cured if you have the operation'. Fred declines the operation, and dies, and the doctor, with no new relevant information, says 'it's very likely that he would have been cured if he had had the operation'. Such pairs could be multiplied indefinitely. For easier arithmetical examples: it's 90% likely that you will get a ball with a black spot if you pick a red ball. It was 90% likely that you would have got a black spot if you had picked a red ball.

I shall now show that standard truth conditions for subjunctives give the wrong answers for uncertain judgements of this form. The argument to follow applies to all accounts of truth conditions which construe a subjunctive conditional as some kind of strict conditional, involving universal quantification over some set of worlds or possibilities, or spelled out in terms of entailment from some premises including the antecedent. I shall stick to the popular Lewis-style truth conditions — roughly, a subjunctive $A \rightarrow C$ is true iff C is true at all closest A-worlds [Lewis 1973; 1979] — see §4 above, though the same points can be made about Goodman-style truth conditions [Goodman 1954] — see §3 above. It also applies to William Lycan [2001], who says a conditional $A \rightarrow C$ is true iff all real and relevant A-possibilities are C-possibilities; and to Bennett [2003], who fine-tunes Lewis's account. Arguably in all of my examples above, and certainly in the last three, the counterfactuals would not come out as highly probable, but as known to be plain false, on these truth conditions. Consider the dog that almost always bites when strangers approach. We can't tell the difference between the cases in which it does and those in which it doesn't. Either there is a bit of indeterminism in play, or it depends on some undetectable subtle feature of the manner of approach. It's not the case, and we take it not to be the case, that in *all* the relevant worlds in which I approached I was bitten. So the truth condition is not satisfied, and we believe it is not satisfied: we think it's certainly false that if you had approached you would have been bitten, according to the truth condition. Similarly for the doctor who thinks it's 90% likely that I will be cured if I have the operation, and later considers whether I would have been cured if I had had the operation. Her uncertainty depends in part on the fine details of what would have happened in the operating theatre. She is certain that the Lewis truth condition does not obtain: that in not all relevant operation-worlds I am cured. Yet she thinks that it's 90% likely that I would have been cured if I had had the operation. And most obviously

of all, the balls in the bag: it is certainly false that in all relevant worlds in which I pick a red ball, it has a black spot. But I say that it's 90% likely that you would have got a black spot if you had picked a red ball.

It might be objected that the probability goes in the consequent of the conditional itself. That is, it's just plain true that (e.g.), if she had had the operation, there would have been a 90% chance of being cured. I have two replies to this objection. First, in all of these examples, it's far from obvious that all relevant antecedent worlds have the same probability of being a *C*-world, or even that in all relevant worlds the consequent has a high chance of being true. The doctor who believes it's 90% likely that I would have been cured, need not believe that that figure would be right for every relevant world in which I go ahead with the operation. Indeed, it is compatible with her belief that she thinks some ways in which the operation could have gone would have had a very low chance of success. Second, even if the first reply is inoperative, it sounds contradictory to say: 'It's certainly not the case that if she had had the operation she would have been cured; but if she had had the operation it is 90% likely that she would have been cured'. That is, there are not really two distinct natural ways of hearing these uncertain conditionals. Scope distinctions are a great philosopher's tool, but we don't naturally hear the two readings of that sentence.

Putting together the point about the close links between wills and woulds, and the above argument about how easy it is for counterfactuals to be plain false on the standard truth conditions, those who run the following combination: conditional probability is the measure of believability of an indicative conditional, Lewis is more or less right about subjunctives — must say that someone may be very confident that you will be cured if you have the operation, or that the dog will bite if you approach, or that you will get a black spot if you pick a red ball; but then, when those have gone counterfactual, but there is no change in their evidence, claim that the corresponding subjunctives are definitely false. This seems to me to be a very unfortunate combination.

A consequence of the above phenomenon is that a very large number of the counterfactuals we accept and assert turn out to be false on the standard truth conditions. Either because of indeterminism, or because determinism is too fine-grained for our everyday antecedents, or because the concepts used in our everyday antecedents don't fit nicely into laws of nature (which play a large role in Lewis's account of closeness, and of course play a crucial role for Goodman), there will be the odd world in which you strike the match and it doesn't light, let alone odd worlds which falsify counterfactuals about human behaviour: 'If you had asked me to do it I would have done so', 'If Fred had been in London he would have got in touch' and so on.[72] (It's important to see that this does not depend on indeter-

[72] Unlike some writers on this theme, I would put less weight on the cases where the probabilities are astronomically high though less than one, which a Lewisian might reasonably ignore; and more weight on cases where the probabilities are, say, around 90%, i.e. significantly different from certainties, i.e. the possibility of error cannot be ignored, yet one does not want to judge the conditional to be certainly false.

minism. Consider the balls-in-bag example as the easiest. Assume determinism. I didn't pick a red ball. So, given the laws and the past, I couldn't have picked a red ball. The supposition that I picked a red ball floats free of the past and the laws. So there is nothing in determinism to say exactly what my hand movements would have been. And not all pickings would have resulted in a black spot.)

I think we can see from these examples that these judgements of 90% probability that C would have happened if A had, are not judgements that it is 90% probable that some fact obtains. What fact? Could God know whether it is TRUE that I would have picked a spotted ball if I had picked a red ball?

I will make a brief remark about Stalnaker's truth conditions. Stalnaker does not treat the conditional as a kind of strict conditional. He says the conditional is true iff the consequent is true at *the* closest A-world. (This has been less popular than Lewis's approach.) Now when the antecedent is false, there is never a unique closest A-world. Think of all the different hand-movements you could have made if you had struck the match or picked a red ball. Stalnaker [1975] adopts the technique of supervaluations to deal with this fact. So what we get is the conditional is true iff the consequent is true at all permissible candidates for closest A-world, i.e. at all closest A-worlds; false iff the consequent is false at all closest A-worlds; otherwise the conditional is indeterminate. Now this is not so uncongenial to the Thesis, which is unsurprising, as Stalnaker's original aim was to find truth conditions compatible with the probabilistic account, and to extend the account to subjunctives. My complaints, transposed to Stalnaker's account, are that vast numbers of subjunctive conditionals just get the verdict 'indeterminate' and this is not very helpful; second, the probability of the *truth* of a conditional is still the same as it is for Lewis, and all my problem cases turn out to be not true. And we do not have a well-developed theory for how to think about how likely it is that if A, C, when it is almost certainly indeterminate. There have been some attempts, by Stalnaker, van Fraassen, Jeffrey, McGee (see §9.4 above), but all ran into difficulties. I am inclined to think that if there were anything promising to be discovered along these lines it would have been discovered by now. But that judgement might be premature.

10.6

Bennett [2003] is aware of the above difficulties, and they get some attention in his book. His first reaction, he calls the near-miss proposal: a subjunctive conditional counts as true iff the consequent is true in almost all the relevant A-worlds. This is equivalent to saying that it's true iff the relevant conditional probability is sufficiently high. This is both vague and context-dependent, but I don't object to that. There are more serious objections to be made. Here are six. First, the proposal allows a conditional to be true which happens to have a true antecedent and false consequent. Bennett amends the account by adding that this is not the case. Second, suppose, just for the sake of argument, that the threshold for truth is around 99%. (I am aware that it is usually vague and context-dependent, but that does not

affect the arguments to follow.) Then if you know that the relevant probability of C given A is 99%, (and you know that the antecedent is false, so the first amendment doesn't apply), you know enough to be sure that the conditional is true, i.e. sure that if A had been the case, C would have been the case (on this account). But you are not: you are just 99% sure that if A had been the case, C would have been the case. Third, if the relevant conditional probability misses the threshold by a small amount, say it is 90%, you should say the conditional is definitely false, and utterly reject 'If A, C'; but you don't, you think it is 90% likely that C would have happened if A had. Fourth, the proposal falls foul of the lottery paradox.[73] As one may put it — probabilities go down on conjunction, but truth values don't! 'If A, B' and 'If A, C' should entail 'If A, then $B\&C$'. Suppose the balls in the bag are numbered 1 to 100. (I pick balls-in-bag examples just to get the structure right.) 'If you had picked a ball, it wouldn't have been number 1' and 'If you had picked a ball it wouldn't have been number 2' can both be true, but 'If you had picked a ball it wouldn't have been number 1 and it wouldn't have been number 2' is false. Fifth, consider conditionals of the form 'If you had tossed the coin n times you would have got at least one heads'. Whatever the threshold, it seems absurd to hold that as n increases there is some value for n at which such conditionals suddenly switch from false to true. Sixth: probabilities change. It may be above the threshold on Monday that if she were to have the operation on Friday she would survive, below the threshold on Tuesday, above again on Wednesday. So the conditional is true on Monday, false on Tuesday, true again on Wednesday. So 'If she had had the operation on Friday, she would have survived' gets different truth values with reference to different times. This is perhaps not a knock-down objection, but we usually think of truth values as more lasting features of our claims than probabilities.

Bennett also considers what he calls a 'more radical proposal: drop truth'; and considers it favourably, which I think is right; but then goes on to say that it doesn't matter very much, and 'does not narrow the theoretical gap between indicatives and subjunctives' (ibid. p. 253). It is odd to say that it doesn't matter very much: without truth, we can no longer think of validity in terms of preservation of truth; we no longer have a ready-made systematic theory of embedded conditionals; Bennett had previously spent some time arguing against the Law of Conditional Excluded Middle, that subjunctives with the same antecedent but contradictory consequents could both be false, and uses that claim subsequently. But without truth, it's not clear what the Law states; and no two such conditionals can each have a conditional probability of less than 50%: their conditional probabilities sum to 1.

What Bennett means is that the careful fine-tuning of the notion of closeness which has occupied many chapters of his book is still needed, whether we go for truth or probability. There is something right about this. All theories of subjunctives — Goodman's, Lewis's, and the probabilistic account — share what is

[73] I owe this point to John Hawthorne [2005].

essentially the same problem, that of specifying what you hang on to, and what you give up, when you make a counterfactual supposition: suppose such-and-such had been the case; what do you hold constant? For Goodman this is the problem of cotenability, for Lewis it is the problem of closeness, for the probabilistic account, it is the problem of which probability distribution is the appropriate one. One can present the probabilistic view in such a way that it is a close relative of Lewis's. First we must specify the class of relevant *A*-worlds. Where Lewis says the counterfactual is true iff *C* is true in all of them, otherwise false, we may alternatively say: take a probability distribution over them, and figure out how likely it is that we have a *C*-world, given that we have an *A*-world. We both have the problem of specifying the class of relevant worlds.

10.7

I used to think that while the probabilistic account could broadly agree with Lewis about what one holds constant, it could do so with minimal fuss. The default time to consider is shortly before the antecedent time. Keep the laws of nature constant. We just need to try to estimate the chance, then, that *C* given *A*.

Unfortunately for all views, things are more complicated than that. We also hold constant independent chance events that occur subsequently, and have some bearing on the consequent.[74] For instance: I'm on my way to the airport. The car breaks down on the motorway. I miss my flight. When the repairman turns up I say 'If I had caught that flight I'd be half way to Paris by now'. 'Which flight were you getting?', he asks. I tell him. 'Well you're wrong', he says, 'It crashed. If you had caught that flight you would be dead by now'.

That's the dramatic example. There's also 'If I had bet on heads I would have won.' 'If I had picked lottery ticket number *xxx*, I would have won'. If the plane was brought down by a rare chance event, very unlikely in advance (so that in advance, the plane was no different in terms of safety from any standard plane), and if my presence or absence on the plane had no causal bearing on whether the crash would occur, it seems, the repairman's remarks are correct, and the rational forward-looking conditional and the rational hindsightful counterfactual come apart. Well, yes, but note that in some sense the person who said in advance for no good reason 'If you catch that plane you will be killed' or 'If you buy ticket *xxx* you will win', and I miss the plane but it crashes, or, I don't buy a ticket but that is the number that comes up, we would say she was right!

So the conditional probability we are interested in, for counterfactuals, and in a sense the ultimate verdict on the forward-looking wills, is the chance, back then, when *A* still had some chance of coming about, of *C* given *A* and any relevant, causally independent, subsequent facts that bear on *C*. You have the chance back then. Then you eliminate the 'no crash' possibilities and consider the probability distribution over the remaining possibilities. It is still a conditional probability, but

[74] I have written about this elsewhere [Edgington 2004].

not one which represents a reasonable degree of conditional belief at the earlier time.

Why do we assess them in this way? Because it is these hindsightful conditionals that feed into our inferential practices, in using them to discover what is the case. Here is an example.

A long time ago, a volcano erupted. It was a slow eruption, the lava creeping slowly forward. When it began, it was very likely that the lava would eventually submerge valley *A*, but valley *B* would not be affected. However, in the unlikely event of an earthquake of a particular kind at an appropriate time, the path of the lava would very probably be switched away from valley *A*, towards valley *B*. As a matter of fact, that is what happened.

Along comes our geologist, centuries later, making her inference about the volcano. She already knows about the earthquake. 'That volcano must have erupted', she concludes. For there is lava in valley *B*. And given what I know about the earthquake, that's what I'd expect to find if that volcano had erupted.'

Suppose there was a second volcano whose potential eruption, at the time in question, presented much more danger to valley *B*, but in the unlikely event of the earthquake, its lava would probably be diverted elsewhere. The hindsightful counterfactuals get things right: if the second volcano had erupted, there would not be lava in valley *B*, and if the first had erupted there would be. These hindsightful judgements stand most chance of leading us to true beliefs. This explains our practice in evaluating these quirky cases.

Or: I'm spotted arriving in Paris several hours late for my appointment. Surprise! 'She must have missed her plane', they say. 'If she had caught that plane she would be dead.'

10.8

When Lewis gave his criteria for closeness in 'Counterfactual Dependence and Time's Arrow', he did so for what he called 'the standard resolution of vagueness' of the similarity relation between worlds. While I prefer the locution 'context-dependence' to 'vagueness', I think he was right in spirit, and would now be inclined to be perhaps more liberal than he was: pretty well any acceptable indicative conditional can 'go counterfactual' in a suitable context. I will take an extreme example (borrowed from Grice), which involves an absolutely minimal ground for an indicative. If the shift to the counterfactual is permissible here, it looks as if it is permissible for any indicative. There is a treasure hunt. The organizer tells me: 'I'll give you a hint: it's either in the attic or the garden'. Trusting the speaker, I think 'If it's not in the attic it's in the garden'. We are competing in pairs: I go to the attic and tip off my partner to search the garden. I discover the treasure. 'Why did you tell me to go to the garden?' she asks. 'Because if it hadn't been in the attic it would have been in the garden: that's what I was told', (or more pedantically: 'that's what I inferred from what I was told'). That doesn't sound wrong in the context. (Maybe the organizer gave someone else a hint: 'It's either in the attic

or the kitchen'. Repeat the scenario, but this player arrived in the attic too late. 'Why did you tell me to go to the kitchen?' asks her partner. 'Because if it hadn't been in the attic it would have been in the kitchen: I inferred that from what I was told'.)

Or consider: 'Why did you hold Smith for questioning?' 'Because we knew the crime was committed by either Jones or Smith — if it hadn't been Jones, it would have been Smith'. There is also a nice example of Van Fraassen's [1981]: the conjuror holds up a penny and claims he got it from the boy's pocket. 'That didn't come from my pocket' says the boy. 'All the coins in my pocket are silver. If that had come from my pocket, it would have been a silver coin'.[75]

This takes off some of the pressure to find *the* account of relevance or closeness. It also allows us to make sense, in context, of 'far out' subjunctives which do not easily fit the standard pattern outlined in the previous section. Nevertheless, the default, most interesting way of assessing subjunctives, feeding into our inferential practices in important ways, is the one sketched there.

10.9

Let us turn to the 'relocation thesis', the thesis that wills and woulds are one kind of conditional, the plain past/present indicatives another. Dudman [1984a, 1984b], was not the first to stress the close relation between 'wills' and 'woulds' (which I think is correct) but he was, I think, the first to draw the conclusion that there are two kinds of conditionals, the 'wills' and 'woulds' are one kind, the plain past and present tense indicatives are another kind.[76] This thesis has been quite influential (for references see p. 130, n. 6, above). Often in the philosophical literature the wills and woulds are treated as something like 'causal conditionals', the others as 'evidential conditionals'. I am against splitting the traditional class of indicative conditionals in this way.

In the traditional class of indicative conditionals, we have a declarative sentence suitable for making a statement, be it about the past, present or future (or indeed timeless), to which a conditional clause is attached, expressing a judgement not categorically but in the context of a supposition; that is, they do essentially the same sort of thing. Ramsey's thesis is plausible for all this class: you are confident in a conditional to the extent that you have a high degree of belief in the consequent on the supposition of the antecedent. Naturally, our grounds tend to be different for statements about the future and statements about the past, and a common and important sort of ground for the 'wills' is that A, if it happens, will cause it to be the case that C. But first, this sort of ground can equally apply to conditionals about the past — 'If she touched that, she got a shock'; and second, it is not the only kind of ground for those about the future: I know the boss told one of his assistants to meet

[75] It is a nice example because Goodman's [1954] paradigm example of a generalization which does not support counterfactuals was 'All the coins in my pocket are silver'.

[76] As mentioned above, Gibbard [1981] independently holds a more moderate version of this view.

me at the station, but I don't know which; so if Bob doesn't come, Ann will come. Even for those that cry out for a causal interpretation, one can tell more or less bizarre non-standard stories. Here is one from Bennett. 'If it rains tomorrow, the roads will be slippery'. But I don't mean that rain will make the roads slippery: the roads are very well constructed and not made slippery by rain. I've just received a leaflet from the council which (a) includes a weather forecast predicting rain; and (b) says they intend to oil the roads tomorrow, warning that this will make the roads slippery. It doesn't look as if it's going to rain, but the council has a first-rate weather forecaster. However, there is some reason to suspect that the leaflet may be a hoax and not genuine. If it rains, that will be evidence that it is genuine, and hence that they will oil the roads, and hence that the roads will be slippery. Of course one would mislead by making that conditional remark without warning that the most obvious ground is not the operative one. But that is pragmatics. No conditional that does not explicitly use causal language like 'produce' or 'make', 'result' or 'outcome' forces a causal reading, though of course it is very often rightly presumed to be asserted on causal grounds. 'If *A* happens, *B* will happen, but *A* won't cause *B* to happen' is never contradictory.

Here are some further points in favour of the traditional distinction: Jackson [1981; 1987] pointed out that while one can believe 'If Oswald hadn't killed Kennedy, things would have been different from the way they actually are', one cannot believe 'If Oswald didn't kill Kennedy, things are different from the way they actually are'. Nor can one believe 'If the Tories win, things will be different from the way they actually will be'. This conforms with the Thesis in that, with indicative conditionals, you are supposing to be the case something taken as an epistemic possibility, and assessing the consequent under that supposition; while the subjunctive may be used when you know that *A* and *C* are actually false, and you make a judgement about what was going to happen, had *A* been true (but actually, didn't happen).

Jackson's point has a converse: we can say non-trivially 'If he had taken arsenic, things would be just as they actually are', whereas it is trivial to say 'If he took arsenic, things are just as they actually are', or 'If the Tories win, things will be just as they actually will be'.

Also, both kinds of indicatives can occur as conditional commands and promises. This is obvious for the 'wills'. For the plain conditionals: 'If he didn't give the lecture, tell the Principal'. 'If you did it, I promise not to tell'; whereas there are no subjunctive commands or promises: 'If he hadn't given the lecture, tell the Principal' and 'If you had done it, I promise not to tell' are nonsense.

Let me now address Dudman's grammatical reason for drawing a new line, the odd behaviour of tenses in the antecedents of conditionals about the future: 'If it rains tomorrow', not 'If it will rain tomorrow'.

The future tense plays (at least) two roles which are often coincident: it indicates that we are speaking about the future; and it indicates that we are making a predication or inference. The two roles can come apart. When something in the future can be taken as a fixed datum, the present tense is natural: 'Term begins on

October 12th', 'The sun sets at 7.03 tomorrow', 'Christmas Day is on a Sunday this year'. Conversely, we use the future tense when we are inferring something about the present, rather than simply observing it. Consider the difference between 'The washing will be dry now' and 'The washing is dry now'. I'd say the latter on feeling it, the former on looking at my watch, and the sky, and figuring that it has been hanging out long enough. Similarly for 'The chicken will be ready now'. (It would be rash to claim that these facts hold in all languages, but they hold in all the dozen-or-so languages I have asked native speakers about.) Even when speaking about the past, the following have different connotations: 'John got home about ten' (I saw him then); 'John will have got home about ten' (I infer that from the fact that he left at nine).

Now, when we make a *supposition* about the future, as I claim we do with 'if': 'If it rains tomorrow, . . . ' 'Suppose England lose tomorrow, . . . ', we are not predicting or inferring rain or defeat. Nor are we supposing that these things are predictable: there is nothing amiss in 'If it rains tomorrow, I'll be very surprised'. We are taking something, hypothetically, for the sake of argument, for granted, as a datum. To make clear that we are not, even hypothetically, in the business of inferring the antecedent, the present tense is in order.

Here are some exceptions from the philosophical literature: 'If she will get the letter tomorrow anyway, we might as well tell her about it today' [Woods, 1997]; 'If Granny will be dead by sundown, we can start selling her clothes right now' (Dudman). The future tense in these antecedents indicates that it is the predictability of the antecedent that is being supposed.

Dudman and his followers have said that the words that follow 'If' in e.g. 'If it rains tomorrow' and 'If England lose tomorrow' do not make a sentence, and thus they have been wrongly construed by philosophers. I disagree. 'It rains tomorrow' and 'England lose tomorrow' are sentences, though, given the nature of the weather and games, not sentences for which we have much use unattached to an 'if'. 'England lose tomorrow' could be used by someone who has fixed the game in advance, or as a statement about what happens in tomorrow's episode of a soap opera he has written. 'It rains at six o'clock' could be said to a newcomer to an equatorial climate where rain is as regular as clockwork, who had been planning his day.

So I think there is an innocent explanation of tense oddity: in 'If it rains tomorrow. . . ' we hypothetically (hence the 'if') take as a datum about the future (hence the present tense) that it rains tomorrow. This syntactic feature does not indicate a distinct kind of conditional thought. It is a consequence of (1) the more general phenomenon of present-tense future reference ('The sun sets at 7.03 tomorrow'); and (2) the nature of suppositions.

11 CONCLUDING REMARKS

None of the main theories of conditionals is incoherent. All are possible ways in which speakers and thinkers could use 'if'. It is an empirical question which theory fits our practice best. Why do philosophers get worked up about it? Why don't we just leave the matter to be settled by questionnaires, or the empirical work of linguists and cognitive psychologists? (And indeed there is much work in this field. Jonathan Evans and David Over [2004] provide an excellent overview.) It is not just an empirical question for philosophers. It is a normative question. We have here an immensely valuable form of thought, without which our thinking would be immeasurably diminished. And we want the theory that best explains why conditionals matter so much to us. As I said earlier, the truth functional theory of indicative conditionals deprives us of the ability to distinguish between believable and unbelievable conditionals whose antecedent we think is unlikely to be true. We would be intellectually impoverished if we used 'if' that way. And I have argued in the last section, a lot of theories of subjunctive conditionals have the consequence that almost all but the most trivial conditionals of this form are knowably false; and this would have a disastrous effect on the use we make of these conditionals. We get worked up because we have the inkling that there is an essential form of thought here, which serves important purposes, and we are after the nature of conditional thinking — an account of how and why this form of thought serves these purposes. That is why there is a philosophy of 'if', but we don't write philosophy books about 'whereas' or 'while'.

Finally, it is worth adding that subjunctive conditionals are supposed to do a lot of work for us within philosophy, as well as in ordinary life. They have been used to 'analyse' causation, dispositions, laws, and play a large part in some accounts of perception and knowledge. On the first, causation, I think we need to appeal to causal notions to get subjunctive conditionals right, and the order of explanation goes that way round. I am a little sceptical about their being a valuable philosophical tool for illuminating other concepts, but I leave that question open.

ACKNOWLEDGEMENTS

Much of the work on the original version of this paper was done while I held a British Academy Research Readership, for which I am very grateful. I am also indebted, for comments and discussion, to Jonathan Bennett, Keith Hossack, David Over, the late Raúl Orayen, David Papineau and Scott Sturgeon. Michael Firestone's thesis, and the manuscript which later became Woods [1997], were also beneficial influences. My greatest philosophical debt is to the work of Ernest Adams.

Dorothy Edgington
Magdalen College, Oxford, UK

BIBLIOGRAPHY

[Adams, 1965] E. W. Adams. A Logic of Conditionals. *Inquiry* **8**, 166–197, 1965.
[Adams, 1966] E. W. Adams. Probability and the logic of conditionals. In Hintikka, J. and Suppes, P. (eds.), 256–316, 1966.
[Adams, 1970] E. W. Adams. Subjunctive and indicative conditionals. *Foundations of Language* **6**, 89–94. 1970.
[Adams, 1975] E. W. Adams. *The Logic of Conditionals*. Dordrecht, Reidel, 1975.
[Adams, 1993] E. W. Adams. On the rightness of certain counterfactuals. *Pacific Philosophical Quarterly* **74**, 1–10, 1993.
[Anderson, 1951] Alan Ross Anderson. A note on subjunctive and counterfactual conditionals. *Analysis* **12**, 35–38, 1951.
[Appiah, 1985] Anthony Appiah. *Assertion and Conditionals*. Cambridge: Cambridge University Press, 1985.
[Appiah, 1986] Anthony Appiah. The importance of triviality. *Philosophical Review* **95**, 209–231, 1986.
[Ayers, 1965] M. R. Ayers. Counterfactual and subjunctive conditionals. *Mind* **74**, 347–364, 1965.
[Bayes, 1940] Thomas Bayes. An essay towards solving a problem in the doctrine of chances, in Deming, W. E. (ed.) 1940. Originally published in *Transactions of the Royal Society of London* **53**, 370–418, 1763. *Philosophical Review* **95**, 209–231, 1986.
[Bennett, 1974] Jonathan Bennett. Review of David Lewis, *Counterfactuals*. *Canadian Journal of Philosophy* **4**, 381–402, 1974.
[Bennett, 1984] Jonathan Bennett. Counterfactuals and temporal direction. *Philosophical Review* **93**, 57–91, 1984.
[Bennett, 1988] Jonathan Bennett. Farewell to the Phlogiston Theory of Conditionals. *Mind* **97**, 509–527, 1988.
[Bennett, 1995] Jonathan Bennett. Classifying conditionals:the traditional way is right. *Mind* **104**, 331–334, 1995.
[Bennett, 2003] Jonathan Bennett. *A Philosophical Guide to Conditionals*. Oxford: Oxford University Press, 2003.
[Bernoulli, 1713] Jacques Bernoulli. *Ars Conjectandi*. Basle, 1713.
[Black, 1950] Max Black. *Philosophical Analysis*. Englewood Cliffs: Prentice Hall, 1950.
[Blackburn, 1986] Simon Blackburn. How can we tell whether a commitment has a truth condition?, in Travis, C. (ed.), 201–232, 1986.
[Burton, 1980] David Burton. *Elementary Number Theory*. Boston:Allyn and Bacon, 1980.
[Carlstrom and Hill, 1978] I. Carlstrom and C. Hill. Review of Adams 1975, *Philosophy of Science* **45**, 155–158, 1978.
[Carnap, 1936] R. Carnap. Testability and meaning. *Philosophy of Science* **3**, 509–527, 1988.
[Chisholm, 1946] R. Chisholm. The contrary-to-fact conditional. *Mind* **55**, 289–307, 1946.
[Davidson, 1980] Donald Davidson. Mental events, in his *Essays on Actions and Events*. Oxford:Clarendon Press, 207–225, 1980.
[Dale, 1985] A. J. Dale. Is the future unreasonable? *Analysis* **45**, 179–183, 1985.
[Deming, 1940] W. E. Deming. *Facsimiles of Two Papers by Bayes*. Washington D.C:US Department of Agriculture, 1940.
[Dudman, 1983] V. H. Dudman. Tense and time in verb clusters of the primary pattern. *Australian Journal of Linguistics* **3**, 25–44, 1983.
[Dudman, 1984] V. H. Dudman. Parsing if-sentences. *Analysis* **44**, 145–153, 1984.
[Dudman, 1984a] V. H. Dudman. Conditional interpretations of 'if-sentences'. *Australian Journal of Linguistics* **4**, 143–204, 1984.
[Dudman, 1986] V. H. Dudman. Antecedents and consequents. *Theoria* **52**, 168–199, 1986.
[Dudman, 1987] V. H. Dudman. Appiah on 'if'. *Analysis* **47**, 74–79, 1987.
[Dudman, 1988] V. H. Dudman. Indicative and subjunctive. *Analysis* **48**, 13–22, 1988.
[Dudman, 1989] V. H. Dudman. Vive la Revolution!. *Mind* **98**, 591–603, 1988.
[Dudman, 1992] V. H. Dudman. Probability and assertion. *Analysis* **52, 4**, 204–211, 1992.

[Dudman, 1994] V. H. Dudman. On conditionals. *Journal of Philosophy* **91**, 113–128, 1994.
[Dummett, 1959] Michael Dummett. Truth, in Dummett, M. 1978, 1–24, 1959.
[Dummett, 1973] Michael Dummett. *Frege:The Philosophy of Language*. London: Duckworth, 1973.
[Dummett, 1978] Michael Dummett. *Truth and Other Enigmas*. London:Duckworth, 1978.
[Dummett, 1992] Michael Dummett. *The Logical Basis of Metaphysics*. London: Duckworth, 1992.
[Edgington, 1986] Dorothy Edgington. Do conditionals have truth conditions?, in Jackson, F. (ed.) 1991, 176–201. First published in *Critica* **18,52**, 3–30, 1986.
[Edgington, 1991] Dorothy Edgington. The mystery of the missing matter of fact. *Proceedings of the Aristotelian Society*, Supplementary Volume *65*, 185–209, 1991.
[Edgington, 2004] Dorothy Edgington. Counterfactuals and the benefit of hindsight. In Phil Dowe and Paul Noordhof, eds., *Cause and Chance*, pp. 12–27. London: Routledge, 2004.
[Eells and Skyrms, 1994] E. Eells and B. Skyrms (eds.)*Probability and Conditionals*. Cambridge:Cambridge University Press, 1994.
[Ellis, 1973] Brian Ellis. The logic of subjective probability. *British Journal for the Philosophy of Science* **24**, 125–152, 1973.
[Ellis, 1979] Brian Ellis. *Rational Belief Systems*. Oxford:Basil Blackwell, 1979.
[Ellis, 1984] Brian Ellis. Two Theories of Indicative Conditionals, 1984. *Australasian Journal of Philosophy* **62**, 50–66.
[Evans and Over, 2004] Jonathan St. B. T. Evans and David E. Over. *If*. Oxford: Oxford Univeristy Press, 2004.
[Fine, 1975] Kit Fine. Critical notice of David Lewis's *Counterfactuals*. *Mind* **84**, 451–58, 1975.
[Firestone, 1995] Michael Firestone. *The Meaning of 'If'. A Study of the Conditional*. MA Thesis, Australian National University, 1995.
[Fowler, 1965] R. W. Fowler. *A Dictionary of Modern English Usage*, second edition, revised by Sir Ernest Gowers. Oxford:Clarendon Press, 1965.
[Frege, 1960] G. Frege. Begriffsschrift, in Geach and Black 1960, 1–20. First published in 1879. 1960.
[Frege, 1979] G. Frege. *Posthumous Writings*. Oxford:Basil Blackwell, 1979.
[Frege, 1980] G. Frege. *Philosophical and Mathematical Correspondence*. Oxford:Basil Blackwell, 1980.
[Geach and Black, 1960] Peter Geach and Max Black. *Translations from the Philosophical Writings of Gottlob Frege*. Oxford:Basil Blackwell, 1960.
[Gibbard, 1981] A. Gibbard. Two recent theories of conditionals in Harper, Stalnaker and Pearce (eds.), pp. 211-247, 1981.
[Gleick, 1987] James Gleick *Chaos*. Penguin Books, 1987.
[Goodman, 1947] N. Goodman. The Problem of Counterfactual Conditionals. *Journal 0f Philosophy* **44**, 113–28, 1947.
[Goodman, 1955] N. Goodman. *Fact, Fiction and Forecast*. Indianapolis: Bobbs-Merrill, 1955.
[Grandy and Warner, 1986] R. E. Grandy and R. Warner (eds.) *Philosophical Grounds of Rationality*. Oxford:Clarendon Press, 1986.
[Grice, 1989] H. P. Grice. *Studies in the Way of Words*. Cambridge MA: Harvard University Press, 1989.
[Hájek, 1989] Alan Hájek. Probabilities of conditionals — revisited. *Journal of Philosophical Logic* **18**, 423–428, 1989.
[Hájek, 1994] Alan Hájek. Triviality on the cheap? In E. Eells and B. Skyrms, eds., *Probability and Conditionals: Belief Revision and Rational Decision*, pp. 113–40. Camb ridge: Cambridge University Press, 1994.
[Harper and Hooker, 1976] W. L. Harper and C. A. Hooker (eds.) *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*, Volume I. Dordrecht:Reidel, 1976.
[Harper, Stalnaker and Pearce, 1981] W. L. Harper, R. Stalnaker and C. T. Pearce (eds.) *Ifs*. Dordrecht:Reidel, 1981.
[Hawthorne, 2005] John Hawthorne. Chance and counterfactuals. *Philosophy and Phenomenological Research* **70**, 396–405, 2005.
[Hintikka and Supper, 1966] J. Hintikka and P. Suppes(eds.) *Aspects of Inductive Logic*. Amsterdam:North Holland, 1966.

[Jackson, 1977] Frank Jackson. A causal theory of counterfactuals. *Australasian Journal of Philosophy* **55**, 13–21, 1977.
[Jackson, 1979] Frank Jackson. On assertion and indicative conditionals. *Philosophical Review* **88**, 565–589, 1979.
[Jackson, 1980] Frank Jackson. Conditionals and possibilia. *Proceedings of the Aristotelian Society*, **81**, 125–137, 1980.
[Jackson, 1987] Frank Jackson. *Conditionals*. Oxford:Basil Blackwell, 1987.
[Jackson, 1990] Frank Jackson. Classifying conditionals. *Analysis* **50**, 134–147, 1990.
[Jackson, 1991] Frank Jackson (ed.) *Conditionals*. Oxford:Clarendon Press, 1991.
[Jeffrey, 1964] Richard Jeffrey. If. *Journal of Philosophy* **61**, 702–703, 1964.
[Jeffrey, 1991] Richard Jeffrey. Matter of fact conditionals. *Proceedings of the Aristotelian Society* Supplementary Volume **65**, 161-183, 1991.
[Kripke, 1963] Saul Kripke. Semantical considerations on modal logic. *Acta Philosophica Fennica* **16**, 83–94, 1963.
[Kripke, 1972] Saul Kripke. *Naming and Necessity*. Oxford:Basil Blackwell, 1972.
[Lance, 1991] Mark Lance. Probabilistic dependence among conditionals. *Philosophical Review* **100**, 269–276, 1991.
[Laplace, 1951] Pierre Simon Laplace. *A Philosophical Essay on Probabilities*. Translated by Truscott, F. W. and Emory, F. L. New York:Dover Publications, 1951.
[Lewis, 1973] David Lewis. *Counterfactuals*. Oxford: Basil Blackwell, 1973.
[Lewis, 1973a] David Lewis. Causation. *Journal of Philosophy* **70**, 556–567, 1973a. Page references to Lewis, 1973a.
[Lewis, 1976] David Lewis. Probabilities of conditionals and conditional probabilities. *Philosophical Review* **85**, 297–315, 1976. Page references to Lewis 1986.
[Lewis, 1979] David Lewis. Counterfactuals and time's arrow. *Nous* **13**, 455–476, 1979. Page references to Lewis 1986.
[Lewis, 1980] David Lewis. A subjectivist's guide to objective chance, in Harper, W.L., Stalnaker, R., and Pearce, C.T. (eds) 1981, 267–297. Page references to Lewis 1986. First published in Jeffrey, R. (ed.) 1980: *Studies in Inductive Logic and Probability*, vol. 2, Berkley and Los Angeles:University of California Press, 263–293, 1980.
[Lewis, 1981] David Lewis. Causal decision theory. *Australasian Journal of Philosophy* **59**, 5–30, 1981. Page references to Lewis 1986.
[Lewis, 1986] David Lewis. *Philosophical Papers* Volume 2. Oxford:Oxford University Press, 1986.
[Lewis, 1986a] David Lewis. Probabilities of conditionals and conditional probabilities II. *Philosophical Review* **5**, 581–9, 1986a.
[Lewis, 1994] David Lewis. Humean supervenience debugged. *Mind*, **103**, 473-490, 1994.
[Lowe, 1990] E. J. Lowe. Conditionals, context and transitivity. *Analysis* **50, 2**, 80–87, 1990.
[Lowe, 1995] E. J. Lowe. The truth about counterfactuals. *Philosophical Quarterly* **43**, 41–59, 1995.
[Lycan, 2001] William Lycan. *Real Conditionals*. Oxford: Oxford University Press, 2001.
[Mackie, 1973] J. Mackie. *Truth, Probability and Paradox*. Oxford:Clarendon Press, 1973.
[McGee, 1985] Vann McGee. A Counterexample to Modus Ponens. *Journal of Philosophy* **82**, 462–71, 1985.
[McGee, 1989] Vann McGee. Conditional probabilities and compounds of conditionals. *Philosophical Review* **98**, 485–542, 1989.
[McGee, 1994] Vann McGee. Learning the impossible, in Eells, E. and Skyrms, B. (eds.) 1994, pp. 179–199, 1994.
[Mellor, 1993] D. H. Mellor. How to believe a conditional. *Journal of Philosophy* **90, 5**, 233–248, 1993.
[Menzies, 1989] Peter Menzies. Probabilistic causation and causal processes: a critique of Lewis. *Philosophy of Science* **56**, 642–663, 1989.
[Pearl, 1988] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems*. San Mateo, California: Morgan Kaufmann, 1988.
[Pendlebury, 1989] Michael Pendlebury. The projection strategy and the truth conditions of conditional statements. *Mind* **390**, 179–205, 1989.

[Quine, 1952] W. V. O. Quine. *Methods of Logic*. London: Routledge and Kegan Paul, 1952. Page references to third edition, 1974.

[Quine, 1966] W. V. O. Quine. On a supposed antinomy, in Quine, W. V. O. Quine, *The Ways of Paradox*. New York:Random House, 21–23. First published in *Mind* **62**, 1953, as On a so-called paradox.

[Ramsey, 1931] Frank Ramsey. *The Foundations of Mathematics*. London: Routledge and Kegan Paul, 1931.

[Read, 1995] Stephen Read. Conditionals and the Ramsey Test. *Proceedings of the Aristotelian Society* Supplementary Volume **69**, 1995.

[Russell, 1919] B. Russell. *Introduction to Mathematical Philosophy*. London: George Allen and Unwin, 1919.

[Russell and Whitehead, 1962] B. Russell and A. N. Whitehead. *Principia Mathematica to *56*. Cambridge: Cambridge University Press, 1962. First published 1910.

[Ryle, 1950] Gilbert Ryle. 'If', 'so' and 'because', in Black (ed.) 1950.

[Sanford, 1989] David H. Sanford. *If P, Then Q: Conditionals and the Foundations of Reasoning*. London: Routledge, 1989.

[Skyrms, 1981] B. Skyrms. The prior propensity account of subjunctive conditionals, in Harper, W. L., Stalnaker, R. and Pearce, G. (eds.), 259–265, 1981.

[Skyrms, 1994] B. Skyrms. Adams conditionals, in Eells, E. and Skyrms, B. (eds.) , 1994, 13–26, 1994.

[Smiley, 1984] Timothy Smiley. Hunter. *Proceedings of the Aristotelian Society* **84**, 113–122, 1984.

[Smith, 1991] Peter Smith. The butterfly effect. *Proceedings of the Aristotelian Society* **91**, 247–267, 1991.

[Stalnaker, 1968] R. Stalnaker. A theory of conditionals in *Studies in Logical Theory*, American Philosophical Quarterly Monograph Series **2**. Oxford: Blackwell, 98–112. Reprinted in Jackson, F. (ed.) 1991. Page references to 1991.

[Stalnaker, 1970] R. Stalnaker. Probability and conditionals. *Philosophy of Science* **37**, 64–80. Reprinted in Harper, W. L., Stalnaker, R, and Pearce, G. (eds.) 1981. Page references to 1981.

[Stalnaker, 1975] R. Stalnaker. Indicative conditionals? *Philosophia* **5**, 269–286. Reprinted in Jackson, F. (ed.) 1991, 136–154. Page references to 1991.

[Stalnaker, 1978] R. Stalnaker. A defense of conditional excluded middle. In Harper *et al.*, eds. pp. 87–104, 1981.

[Stalnaker, 1981] R. Stalnaker. A defense of conditional excluded middle, in Harper, W. L., Stalnaker, R, and Pearce, G. (eds.) 1981. Page references to 1981.

[Stalnaker, 1984] R. Stalnaker. *Inquiry*. Cambridge MA: MIT Press, 1984.

[Stalnaker, 2005] R. Stalnaker. Conditional propositions and conditional assertions. In *New Work on Modality: MIT Working Papers in Linguistics and Philosophy*, 51, 2005.

[Stalnaker and Jeffrey, 1994] R. Stalnaker and R. Jeffrey. Conditionals as random variables, in Eells, E. and Skyrms, B. (eds.), 31–46, 1994.

[Strawson, 1986] P. F. Strawson. 'If' and '$\supset$', in Grandy, R. E. and Warner R., 229–242, 1986.

[Thomson, 1990] James Thomson. In defense of $\supset$. *Journal of Philosophy* **87**, 56–70, 1990.

[Tichy, 1976] Pavel Tichy. A counterexample to the Stalnaker–Lewis analysis of counterfactuals. *Philosophical Studies* **29**, 271–273, 1976.

[Travis, 1986] Charles Travis (ed.) *Meaning and Interpretation*. Oxford:Basil Blackwell, 1986.

[van Fraassen, 1976] Bas Van Fraassen. Probabilities of conditionals, in Harper, W. and Hooker, C. (eds.), 261–308, 1976.

[van Fraassen, 1980] Bas Van Fraassen. Review of Ellis 1979. *Canadian Journal of Philosophy* **10**, 497–511, 1980.

[van Fraassen, 1981] Bas van Fraassen. Essences and laws of Nature. In R. Healey, ed., *Reduction, Time and Reality*, pp. 189–200. Cambridge: Cambridge University Press, 1981.

[Von Wright, 1957] G. H. Von Wright. *Logical Studies*. London: Routledge and Kegan Paul, 1957.

[Woods, 1997] Michael Woods. *Conditionals*. Oxford: Oxford University Press, 1997.

[Wright, 1983] Crispin Wright. Keeping track of Nozick. *Analysis* **43**, 134–140, 1983.

[Wright, 1992] Crispin Wright. *Truth and objectivity*. Cambridge MA: Harvard University Press, 1992.

DAG WESTERSTÅHL

QUANTIFIERS IN FORMAL AND NATURAL LANGUAGES

For a long time, the word 'quantifier' in linguistics and philosophy simply stood for the universal and existential quantifiers of standard predicate logic. In fact, this use is still prevalent in elementary textbooks. It seems fair to say that the dominance of predicate logic in these fields has obscured the fact that the quantifier expressions form a *syntactic category*, with characteristic interpretations, and with many more members than $\forall$ and $\exists$.

Actually, when Frege discovered predicate logic, it was clear to him that the universal and existential quantifiers were but two instances of a general notion (which he called *second level concept*). That insight, however, was not preserved during the early development of modern logic. It took quite some time before the mathematical machinery behind quantification received, once more, an adequate genera formulation. This time, the notion was called *generalised quantifier*; a first version of it was introduced by Mostowski in the late 1950s. Logicians gradually realised that generalised quantifiers were an extremely versatile syntactic and semantic tool — practically anything one would ever want to say in any logic can be expressed with them. The power of expression, properties and interrelations of various logics with generalised quantifies is now a well established domain of study in mathematical logic.

This is the mathematical side of the coin. The linguistic side looks a bit different. Syntactically, there are many expressions one could place in the same category as *some* and *every*: *no, most, many, at least five, exactly seven, all but three*, ... These expressions — the *determiners* — occur in *noun phrases*, which in turn occur as subjects, objects, etc. in the *NP–VP* analysis of sentences usually preferred by linguists. Logically, however, subject–predicate form had fallen into disrepute since the breakthrough of predicate logic. So it was not obvious how to impose a semantic interpretation on these syntactic forms — except by somehow rewriting them in predicate logic. This may explain why the systematic study of quantifiers in natural language is of a much later date than the one for mathematical language.

The starting-point of this study was when Montague showed that linguistic syntax is, after all, no insurmountable obstacle to systematic and rigorous semantics. Montague did not yet have the quantifiers in a separate category. But in 1981 Barwise and Cooper united Montague's insights with the work on generalised quantifiers in mathematical logic in a study of the characteristics of natural language quantification [Barwise and Cooper, 1981]. At about the same time, but independently and from a slightly different perspective, Keenan and Stavi were investigating the semantic properties of determiner interpretations [Keenan and Stavi, 1986]. It became clear that, in natural language too, the quantifier category is quite rich and semantically powerful. In the few years that have passed since then, the

D.M. Gabbay and F. Guenthner (eds.), Handbook of Philosophical Logic, 2nd Edition, Volume 14, 223–338.

subject has developed considerably. In particular, van Benthem has discovered an interesting logical theory behind the mechanisms of natural language quantification — often with no direct counterpart for mathematical language [van Benthem, 1984a].

My main aim in this chapter is to give a comprehensive survey of the logic and semantics of natural language quantification, concentrating on the developments in the last five years or so. the basic tools are the generalised quantifiers from mathematical logic. but it is the *questions* asked about quantifiers, not the methods used, that distinguishes our present perspective on quantifiers from that of mathematical logic.

The basic question facing anyone who studies natural language quantification from a semantic viewpoint can be formulated as follows. Logically, the category of quantifiers is extremely rich. For example, even on a universe with *two* elements, there are $2^{16} = 65536$ possible (binary) quantifiers (the reader who finds this hard to believe may wish to turn directly to Section 4.6 for the explanation). But, in natural languages, just a small portion of these are 'realised' (512, according to Keenan and Stavi). Which ones, and why? What are the *constraints* on determiner interpretations in natural language? what are the *properties* of quantifiers satisfying those constraints.

Most of this paper presents various answers to such questions. But we start, in Section 1, with a selective history of quantifiers: from Aristotle via Frege to modern generalised quantifiers. It will be seen that both Aristotle's and Frege's contributions compare interestingly to the recent developments. That section also gives a thorough introduction to generalised quantifiers, and to some logical issues pertaining to them. In particular, the *logical expressive power* of *monadic* quantifiers is discussed in some detail. Section 2 presents basic ideas of the Montague–Barwise–Cooper–Keenan–Stavi approach to natural language quantification. A number of examples of English quantifier expressions are also collected, as empirical data for alter use. In Section 3, several constraints on quantifiers are formulated and discussed and various properties of quantifiers are introduced. The constraints can also be seen as potential *semantic universals*. Section 4 then presents various results in the *theory* of quantifiers satisfying certain basic constraints; results on how to classify them under various aspects, on how to represent them, on their inferential behaviour and other properties. The paper ends with a brief further outlook and two appendices, one on branching quantification and the other on quantifiers as variables.

This chapter is concerned with the *semantics* of quantification. It examines certain precisely delimited classes of quantifiers that arise naturally in the context of natural language. These classes are related in various ways to the (loosely delimited) class of *natural language quantifiers*, i.e. those that are denotations of natural language determiners. I will make few definite claims about the exact nature of this relationship, but I will discuss several tentative proposals. The idea is to present the *possibilities* for determiner interpretation, and to give a framework sufficiently general for serious discussion of natural language quantifiers, yet re-

stricted in significant ways compared with the generalised quantifier framework of mathematical logic. (I also hope to make it clear that interesting logical issues arise in the restricted framework (and sometimes only in that framework), and thus that logic can fruitfully be inspired by natural language as well as by the language of mathematics.)

So, except for a few rather straightforward things, I shall have little to say about the *syntax* of quantification here. And except for an introductory overview, I will not attempt to survey generalised quantifiers in mathematical logic. For more on quantification and linguistic theory, cf. [Cooper, 1983] or [van Eijck, 1985]. A very comprehensive survey of quantifiers in mathematical logic is given in [Barwise and Feferman, 1985].

The semantic framework used here is that of classical model theory. It is simple, elegant and well known. that it works so well for natural language quantification too is perhaps a bit surprising. However, there are certain things it does not pretend to handle, for example, intensional phenomena, vagueness, collectives, or mass terms. So these subjects will not be taken up here. but then, they receive ample treatment in other parts of this Handbook.

The logical techniques we need are usually quite elementary. the reader should be used to logical and set-theoretic terminology, but, except on a few occasions, there are no other specific prerequisites (the chapter by Hodges in this Handbook gives a suitable background; occasionally, part of the chapter by van Benthem and Doets will be useful). I have intended to make the exposition largely self-contained, in the sense that (a) most proofs and arguments are given explicitly, and (b) when they are not given, references are provided, but he reader should be able to get a feeling for what is going on without going to the references. Naturally, if these intentions turn out not to be realised, it does not follow that the fault lies with the reader.

This is a survey, and most results are from the literature, although several are new, or generalised, or proved differently here. I have tried to give reasonable credit for known results.

1 BACKGROUND FROM ARISTOTELIAN TO GENERALISED QUANTIFIERS

This section gives a condensed account of the development of what can be called *the relational view of quantifiers*. As a chapter in the history of logic, it seems not to be widely known, which is why I have included a subsection on Aristotle and a subsection on Frege. My main purpose, however, is to introduce a precise concept of quantifier sufficiently general to serve as a basis for what will follow. This is the notion of a *generalised quantifier* from mathematical logic. In the last subsections, I will also mention some of the things mathematical logicians do with quantifiers, as a background to what linguistically minded logicians might do with them.

1.1 Aristotle

Aristotle's theory of syllogisms, for ages considered the final system of logic, is not often seen as rather pointless formal exercise, whose main achievement is to have hampered the development of logic for some two thousand years. But to understand Aristotle's contribution to logic one must distinguish his views from those of his followers. It is a fact that most of his followers and commentators were unable, for various reasons, to appreciate his logical insights (to take one simple but important example the point of using *variables*).[1] From the standpoint of modern logic, on the other hand, these insights ought to be easily visible.

There is, however, one obscuring issue. According to widespread opinion, the breakthrough of modern logic rests upon the *rejection* of a basic Aristotelian idea, namely, that sentences have *subject-predicate form*. This was Russell's view, apparently vindicated by the absence of subject-predicate form in today's standard predicate logic. Hence, Aristotle's logic seems to be built on a fundamental mistake.

If we set aside questions concerning the historical causes of the long standstill in logic after Aristotle, there is, however, no necessary incompatibility between modern logic and subject-predicate form.[2] It is quite feasible to give an adequate account of both relations and quantification while preserving subject-predicate form, as we shall see in 2.3. Thus, although it is true that Aristotle's logic could not adequately account for these things, and thus was unable to express many common forms of reasoning, this weakness is not necessarily tied to his use of subject-predicate form.

In addition to matters of syntactic form, however, one ought to consider the *concepts* Aristotle introduced with his logic, the *questions* he raised about it, and the *methods* he used to answer them. Herein lies his greatest contribution.

Thousands of pages have been written on Aristotle's logic, most of them about irrelevant and futile matters (such as the order between the premisses in a syllogism, why he didn't mention the fourth figure, whether a valid syllogism can have a false premiss — Aristotle himself had no doubts about this — , etc.). Readable modern expositions, with references to the older literature, are Łukasiewicz [1957] and Patzig [1959]. Below I wish to point, without (serious) exegetic pretensions, to one important aspect of Aristotle's logic.

The syllogistics is basically a theory of *inference patterns among quantified sentences*. Here a quantified sentence has the form

(1) QXY,

[1] Actually, contemporaries of Aristotle, like Theophrastus, seem to have understood him rather well. But the medieval reintroduction of Aristotle's logic lost track of many important points. Even 19th century commentators continue in the medieval vein; cf. [Łukasiewicz, 1957].

[2] About the historical causes Russell may well be right. Note that we are also setting aside here the metaphysical claims of Russell's logical atomism, according to which the logical form of sentences mirror the structure of reality.

where X, Y are *universal terms* (roughly 1-place predicate) and Q is one of the quantifiers *all, some, no, not all*. In practice, Aristotle treated these quantifiers as *relations* between the universal terms.[3]

Aristotle chose to study a particular type of inference pattern with sentences of the form (1), the syllogisms. A *syllogism* has two premisses, one conclusion, and three universal terms (variables). Each sentence has two different terms, all three terms occur in the premisses, and one term the 'middle' one, occurs in both premisses but not in the conclusion. It follows that the syllogisms can be grouped into four different 'figures', according to the possible configurations of variables:

$$\begin{array}{cccc} Q_1Zy & Q_1YZ & Q_1ZY & Q_1YZ \\ \underline{Q_2XZ} & \underline{Q_2XZ} & \underline{Q_2ZX} & \underline{Q_2ZX} \\ Q_3XY & Q_3XY & Q_3XY & Q_3XY \end{array}$$

Here the Q_i can be chosen among the above quantifiers, so there are $4^4 = 256$ syllogisms. As a matter of historical fact, Aristotle's specification of the syllogistic form was not quite accurate; he had problems with defining the middle term, and his systematic exposition does not mention the fourth figure (although he in practice admitted syllogisms of this form), but these are minor defects.

Now, the question Aristotle posed — and, in essence, completely answered — can be formulated as follows:

For what choices of quantifiers are the above figures valid?

For example, of we in the first figure let $Q_1 = Q_2 = Q_3 =$ *all*, a valid syllogism results ('Barbara', in the medieval mnemonic); likewise if $Q_1 = Q_2 =$ *no* and $Q_2 =$ *all* ('Celarent'). Note that Aristotle's notion of validity is essentially the modern one: a syllogism is valid if each instantiation of X, Y, Z verifying the premisses also verifies the conclusion (a slight difference is that Aristotle didn't allow the empty or the universal instantiation; this can be ignored here).

There are interesting variants of this type of question. Given some common quantifiers, we can ask for their inference patterns, and try to systematise the answer in some perspicuous way (axiomatically, for example). This is a standard procedure in logic. But we can also turn the question around and ask which quantifiers satisfy the patterns we found: only the ones we started with or others as well? If our common schemes of inference *characterise* our common quantifiers, we have one kind of explanation of the privileged status of the corresponding 'logical constants', and one goal of a theory of quantifiers has been attained.

The latter question is somewhat trivialised in Aristotle's framework, since there were only four quantifiers. For example, the question of which quantifiers satisfy the scheme:

$$\begin{array}{c} QZY \\ \underline{QXZ} \\ QXY \end{array}$$

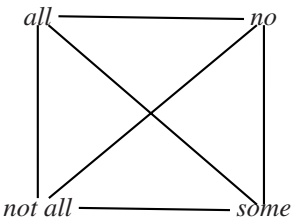
[3]He sometimes comes very close to an explicit statement; cf. the last pages of [Patzig, 1959].

has the obvious answer: just *all*. But the question itself does not depend on the quantifier concept you happen to use. In 4.1 we shall return to it (and in 4.5 to the characterisation of our most common quantifiers), this time with infinitely many quantifiers to choose from, and find some non-trivial answers.

Thus, not only did Aristotle introduce the relational concept of quantifiers, he also asked interesting questions about it. His methods of answering these questions were *axiomatic* (for example, he derived all valid syllogisms from the two syllogisms 'Barbara' and 'Celarent' mentioned above) as well as *model-theoretic* (non-validity was established by means of counter-examples). Even from a modern point of view, his solution leaves only some polishing of detail to be desired. Perhaps this finality of his logic was its greatest 'fault'; it did not encourage applications of the new methods to, say, other inference patterns. Instead, his followers managed to make a sterile church out of his system, forcing logic students to rehearse syllogisms far into our own century. But we can hardly blame Aristotle for that.

It should be noted that outside of logic Aristotle studied quantifiers without restriction to syllogistic form. For example, he made interesting observations on sentences combining negation and quantification (cf. [Geach, 1972]).

We shall not pursue the fate of the relational view of quantifiers between Aristotle and Frege. Medieval logicians spent much time analysing quantified sentences, but they were more or less prevented from having a *concept* of quantifier by their insistence that quantifier words are *syncategorematic*, without independent meaning (this view, incidentally, is still common). Later logicians applied the mathematical theory of relations (converses, relative products, etc.) to give explicit formulations of Aristotle's relational concept, and to facilitate the proofs of his results on syllogisms (cf. [DeMorgan, 1847] or, for a more recent account [Lorenzen, 1958]). These methods were in general only applied to the quantifiers in the traditional *square of opposition* and their converses. A systematic study of quantifiers as binary relations did not appear until the 1980s (cf. Section 4.1).



1.2 Frege

It is undisputed that Frege is the father of modern logic. He invented the language of predicate calculus, and the concept of a formal system with syntactic formation and inference rules. Moreover, his work was characterised by an exceptional theoretical clarity, greatly surpassing that of his contemporaries, and for a long time

also his successors, in logic.

There is some difference of opinion, however, as to how 'modern' Frege's conception of logic was. According to Dummett [1973; 1981], we find in Frege, implicitly if not explicitly, just that dualism between a syntactic (proof-theoretic) and a semantic (model-theoretic) viewpoint which is characteristic of modern logic. "Frege would have had within his grasp the concepts necessary to frame the notion of completeness of a formalisation of logic as well as its soundness" [Dummett, 1981, p. 82] Dummett also traces the notion of an interpretation of a sentence, and thereby the semantic notion of logical consequence, to Frege's work.

This evaluation is challenged in [Goldfarb, 1979], a paper on the quantifier linearly (modern) logic. Goldfarb holds the notion of an interpretation to be nonexistent in Frege's logic: first, because there are no non-logical symbols to interpret, and second, because the universe is fixed once and for all. The quantifiers range over this universe, and the laws of logic are *about* its objects. furthermore, the logicism of Frege and Russell prevented them, according to Goldfarb, from raising any metalogical questions at all.

Although it takes us a bit beyond a mere presentation of Frege's notion of quantifier, it is worthwhile trying to get clear about this issue. The main point to be made is, I think, that Frege was the only one of the logicians at the time who maintained a sharp distinction between syntax and semantics, i.e. between the expressions themselves and their denotations. This fact alone puts certain metalogical questions 'within the reach' of Frege that would have been meaningless to others. Thus, one cannot treat Frege and Russell on a par here. Moreover, if one loses sight of this, one is also likely to miss the remarkable fact that, while the invention of predicate logic with the universal and existential quantifiers can also be attributed to Peano and Russell, Frege was the only one who had a mathematically precise *concept* of quantifier. This concept seems indeed to have gone largely unnoticed among logicians, at least until the last decade or so; in particular, the inventors of the modern generalised quantifiers do not seem to have been aware of it.

For this reason, Frege, but not Russell, has a prominent place in an historical overview of the relational view of quantifiers — in fact, Russell's explanations of the meaning of the quantifiers are in general quite bewildering (for example, [Russell, 1903, Chapter IV, Sections 59–65], or [Russell, 1956, pp. 64–75 and 230–231]). I will present Frege's concept below, and then return briefly to the issue of how questions of soundness and completeness relate to Frege's logic.

1.2.1 Quantifiers as second level concepts

Let us first recall some familiar facts about Frege's theoretical framework.[4] All entities are either *objects* or *functions*. These categories are primitive and cannot be defined. Functions, however, are distinguished from objects in that they have

[4]More precisely, the system of *Grundgesetze* [1893]. The English translation of the first part of this work by M. Furth is prefaced with an excellent introduction, where more details about Frege's conceptual framework can be found.

(one or more) empty places (they are 'unsaturated'). When the empty places are 'filled' with appropriate *arguments* a *value* is obtained. The value is always an object, but the arguments can either be objects, in the case of *first level* functions, or other functions: *second level* functions take first level functions as arguments, etc. — no mixing of levels is permitted. All functions are total (defined for all arguments of the right sort). They can be called *unary, binary*, etc. depending on the number of arguments.[5]

Concepts are identified with functions whose values are among the two truth values *True* and *False*. Thus they have levels and 'arities' just as other functions.

The meaningful expressions in a logical language ('Begriffsschrift') are simple or complex *names* standing for objects or functions.[6] Names have both a sense ('Sinn') and a denotation ('Bedeutung'); only the denotation matters here. there is a strong parallelism between the syntactic and the semantic level: function names also have empty places (marked by special letters) that can (literally) be filled with appropriate object or function names. In particular, sentences are (complex) object names, denoting truth values.

Complex function names can be obtained from complex object names by deleting simple names, leaving corresponding empty places. For example, from the sentence

$$23 \text{ is greater than } 14$$

we obtain the first level function (concept) names

$$x \text{ is greater than } 14,$$
$$23 \text{ is greater than } y,$$
$$x \text{ is greater than } y,$$

and also the second level

$$\Psi(23, 14).$$

Now, suppose the expression

(1) $F(x)$

is a unary first level concept name. Then the following is a sentence.[7]

(2) $\forall x F(x)$.

[5]This notion of 'arity' does not tell us the number of arguments of the arguments, etc; for levels grater than one; we will not need that here.

[6]Actually, Frege did not use "name" for expressions referring to functions. Instead he used "incomplete expression" and the like.

[7]Here I depart from Frege by (i) using modern quantifier notation, and (ii) using the same letter 'x' in (1) and (2). According to Frege, the variable in (1) just marks a place and does not really belong to the concept name, whereas in (2) it is an inseparable part of a function name (cf. below). These distinctions, while interesting, are not essential in the present context.

According to Frege, (2) is obtained by inserting the concept name (1) as an argument in the *second level concept name*

(3) $\forall x \Psi(x)$.

(3) is a *simple* name in Frege's logic. It denotes a unary second level concept, namely, the function which, when applied to any unary first level function $f(x)$, gives the value *True* if $f(x)$ has the value *True* for all its arguments, *False* otherwise.[8]

This, of course, is a version of the usual truth condition for universally quantified sentences: (2) is true iff $F(x)$ is true for all objects x. But Frege's formulation makes it clear that (3) denotes just one of many possible second level concepts, for example,

(4) $\neg\forall x \neg\Psi(x)$

(5) $\forall x(\Phi(x) \rightarrow \Psi(x))$.

(4) is the existential quantifier. (5) is the *binary* second level concept of *subordination* between two unary first level concepts. Both can be defined by means of (3) in Frege's logic, and are thus denoted by complex names.

In a similar fashion, quantification over first level functions can be introduced by means of third level concepts, and so on.

Summarising, we find that there is a well defined *Fregean concept of quantifier*:

> *Syntactically, (simple) quantifier names can be seen as variable-binding operators (but see Note 7 on Frege's use of variables). Semantically, quantifiers are second level concepts.*

If we let, in a somewhat un-Fregean way, the *extension of an n-ary first level concept* be the class of n-tuples of objects falling under it, and the *extension of an n-ary second level concept* the class of n-tuples of extensions of first level concepts falling under it, then the extensions of the quantifiers (3)–(5) are

(6) $\forall_u = \{X \subseteq U : X = U\}$

(7) $\exists_u = \{X \subseteq U : X \neq \varnothing\}$

(8) $\mathbf{all}_i = \{\langle X, Y\rangle : X \subseteq U \& Y \subseteq U \& Y \subseteq Y\}$,

where U is the class of all objects. Apart from the fact that the universe is fixed here (and too big to be an element of a class), these extensions are generalised quantifiers in the model-theoretic sense; cf. Section 1.4.

[8] Note that the quantifier (3) must be defined for all unary first level functions (not only for concepts), since functions are total. As we can see, $\forall x \Psi(x)$ is *false* for arguments that ar enot concepts.

1.2.2 Unary vs. binary quantifiers

Frege was well aware that the usual quantifier words in natural language stand for *binary* quantifiers. For example, in 'On Concept and Object' he writes

> ... the words 'all', 'any', 'no', 'some' are prefixed to concept-words. In universal and particular affirmative and negative sentences, we are expressing *relations between concepts*; we use the words to indicate a special kind of relation ([Frege, 1892, p. 48], my italics).

But he also found that these binary (Aristotelian) quantifiers could be defined by means of the unary (3) and sentential connectives. This was no trivial discovery at the time, and Frege must have been struck by the power and simplicity of the unary universal quantifier. In his logical language he always chose it as the sole primitive quantifier.

The use of unary quantifiers was to become a characteristic of predicate logic, and the success of formalising mathematical reasoning in this logic can certainly be said to have vindicated Frege's choice. It does not follow from this, however, that the same choice is adequate for formalising natural language reasoning. Indeed, we will see later that unary quantifiers are unsuitable as denotations of the usual quantifier words, and that, furthermore, it is simply not the case that all binary natural language quantifiers can be defined by means of unary ones and sentential connectives.

Such a preference for binary quantifiers in a natural language context is, as we can see from the foregoing, in no way inconsistent with Frege's view on quantifiers.[9]

1.2.3 Logical truth and metalogic

Let us return to the DummettGoldfarb dispute about whether metalogical issues such as completeness were in principle available to Frege. The usual notion of completeness of a logic presupposes the notion of logical truth (or consequence),

[9]There may be deeper reasons for preferring binary quantifiers. For example, [Dummett, 1981] regards Frege's decision to use a unary quantifier as *the* fatal step which eventually led to paradox in his system. This is because in the unary case we quantify over all objects, whereas binary quantifiers can restrict the domain to that part of the universe denoted by the first argument (as we will see in Section 2), thereby avoiding the need to consider a total universe [Dummett, 1981, p. 227].

This argument may point to one cause of Frege's actual choice of an inconsistent system, but it is not by itself conclusive against unary quantifiers. The lesson of the paradoxes is not necessarily that one must not quantify over all objects. Indeed, the Tarskian account of the truth conditions for universally quantified sentences is quite independent of the size of the universe, and logicians often quantify over total domains, e.g. the domain of all sets in Zermelo–Fraenkel set theory, without fearing paradox. (It is another matter that they,for 'practical' reasons, often prefer set domains when this is possible.) So the above argument can only have force, I think, when combined with a general theory of meaning of the type that Dummett advocates (and which in some sense rejects the Tarskian account). These deeper issues in the theory of meaning will not be discussed here.

i.e. truth in all models. But the latter notion was clearly not considered by Frege. As Goldfarb remarks, he had no non-logical constants whose interpretation could vary (it seems that he explicitly rejected the use of such constants; cf. Hodges' chapter, section 17), nor did he consider the idea that the universe could be varied. One universe was enough, namely, the universe U of all objects, and only simple truth in U interested Frege.

However, the notion of truth in U is very close to the notion of logical truth. to fix ideas, consider some standard version of *higher-order logic* (say, the logic L_ω presented in the chapter by van Benthem and Doets, Section 3.1). For the purposes of the present discussion we may identify *Frege's logic* with higher-order logic *without* non-logical symbols.[10] Then we can observe that Frege did not 'miss' any standard logical truths. For, each sentence ψ in L_ω has an obvious translation ψ^* in Frege's logic, obtained by 'quantifying out' the non-logical constants. For example,

$$\forall x Px \to Pa$$

translates as

$$\forall X \forall y (\forall x Xx \to Xy),$$

and similarly for higher-order sentences. It is evident that

(9) if ψ is logically true then ψ^* is true in U.

A parenthetical observation is necessary here. Logical truth is often defined as truth in all *set* models, instead of truth in *all* models, whether sets or not. The latter notion is *real* logical truth, and it is with respect to this notion that (9) is evident. As Kreisel has stressed, use of the former notion is only justified for first-order logic, since there the two notions coincide (this follows from the usual completeness proofs). For higher-order sentences, on the other hand, this is open; cf. [Kreisel, 1967].

For first-order logic, there is a converse to (9), provided we disregard sentences such as

$$\exists x \exists y (x \neq y),$$

which have *finite* counter-examples but are still true in the infinite U:

THEOREM 1. *Let M be any infinite class and ψ a first-order sentence. then ψ is true in all infinite models iff ψ^* is true in M.*

[10] Frege's logic, that is, not his whole system with its (inconsistent) principles of set existence (abstraction). The proposed identification slurs over some details, but is consistent with Frege's idea that logic is about a domain of *objects* (U), upon which a structure of functions of different levels is built, with no mixing between functions and objects, or between functions of different levels.

Proof. (This proof uses some standard techniques of first-order model theory; they can be found in [Chang and Keisler, 1973];; but will not be employed in the sequel.) From left to right this is similar to (9); if only set models are considered we employ Kreisel's observation mentioned above. For the other direction, suppose that $\neg\psi = \neg\psi(P, \ldots)$ has an infinite model $\mathbf{N} = \langle N, R, \ldots\rangle$. Again by Kreisel's observation, we can assume that N is a set. Now distinguish two cases, depending on whether M is a set or not. If M is a set, application of the Löwenheim–Skolem–Tarksi theorem gives us a model $\mathbf{M}_0$ of $\neg\psi$ with the same cardinality as M. Via a bijection from M to $M_0, \mathbf{M}_0$ is isomorphic to a model $\langle M, S, \ldots\rangle$ of $\neg\psi$ with universe M. Thus, $\exists X \ldots \neg\psi(x, \ldots)$ is true in M, i.e. ψ^* is false in M, as was to be proved. Now suppose M is a proper class. Starting with $\mathbf{N} = \mathbf{N}_0$ as before, define uniformly for each ordinal α a model $\mathbf{N}_\alpha$ such that $\mathbf{N}_\alpha$ is a proper elementary extension of $\mathbf{N}_\beta$ when $\beta < \alpha$. The union $\mathbf{M}'$ of all these is then a model of $\neg\psi$ (Tarski's union lemma). Moreover, M' is a proper class, whence there is a bijection from M to M'. It follows as before that ψ^* is false in M. ■

Thus, in a sense it makes no difference for first-order logic if we have, as Frege did, a fixed infinite universe (such as U) and no non-logical constants. More precisely, it follows from the above that the true Π^1_1 sentences of Frege's logic correspond exactly to the standard first-order logical truths on infinite models.

In conclusion, then, we have seen that notions such as completeness and soundness were not directly available to Frege, since they presuppose a notion of logical truth he did not have. But Dummett's position is still essentially correct, I think: Frege's work does introduce a version of the dualism between model theory and proof theory. For, Frege had the notion of *truth*, which he certainly did not confound with *provability*. Clearly he considered all theorems of his system to be true. He did not, as far as we know, raise the converse question of whether all true sentences are provable, but surely it was 'within his grasp'. And for his *logic*, this question turns out to be a version of the completeness question, as noted above. Moreover, the answer is *yes* if we restrict attention to Π^1_1 sentences (by the above result and the completeness of first-order logic), *no* otherwise (higher-order logic is not complete).

1.3 Mostowskian Quantifiers

As we know, Frege's work was neglected in the early phase of modern logic, and the rigor he attained, especially in semantics, was not matched for a long time. But the language of predicate logic was powerful enough to be a success even in the absence of a solid semantic basis. In the history of quantifiers, this period is mainly interesting for its discussions on the role of quantification over infinite domains for the foundation of mathematics, but that is not a subject here.

The idea of a mathematically sharp dividing line between syntax and semantics began to reappear gradually in the 1920s, but not until Tarski's truth definition in 1936 did the notion of truth (in a model) become respectable. Tarski's truth con-

ditions for universally and existentially quantified formulas treat $\forall$ and $\exists$ syncategorematically, but it is natural to try some other quantifiers here, i.e. to consider formulas

$$Qx\psi$$

for Q other than $\forall$ and $\exists$. For example, it is clear what the truth conditions for $\exists_{\geq n}$ and $\exists_{=n}$ should look like. To get a general concept, however, we must treat Q non-syncategorematically, i.e. we must have a *syntactic category* 'quantifier' with a specified range of interpretations. Such a general concept appeared in [Mostowski, 1957].

Recall that Tarski defines the relation

$$\mathbf{M} \vdash \phi[g],$$

('g satisfies ϕ in $\mathbf{M}$'), where $\mathbf{M}$ is model, g an assignment of elements in M to the variables, and ϕ a formula. When ϕ is $\forall x\psi$ or $\exists x\psi$, this can be expressed as a condition on the *set*

$$\psi^{\mathbf{M},g,x} = \{a \in M : \mathbf{M} \vDash \psi[g(a/x)]\}.$$

Thus,

$$\mathbf{M} \vDash \forall x\psi[g] \Leftrightarrow \psi^{\mathbf{M},g,x} = M,$$
$$\mathbf{M} \vDash \exists x\psi[g] \Leftrightarrow \psi^{\mathbf{M},g,x} \neq \varnothing,$$
$$\mathbf{M} \vDash \exists_{\geq n} x\psi[g] \Leftrightarrow| \ \psi^{\mathbf{M},g,x} \ |\geqq n.$$

A condition on subsets of M is, extensionally, just a set of subsets of M. So Mostowski defines a (local) *quantifier on* M to be a set of subsets of M, whereas a (global) *quantifier* is a function(al) $\mathbf{Q}$ assigning to each non-empty set M a quantifier $\mathbf{Q}_M$ on M. Syntactically, a quantifier symbol Q belongs to $\mathbf{Q}$, such that $Qx\psi$ is a formula whenever x is a variable and ψ is a formula, with the truth condition

$$\mathbf{M} \vDash Qx\psi[g] \Leftrightarrow \psi^{\mathbf{M},g,x} \in \mathbf{Q}_M.$$

Examples of such quantifiers are

$$\forall_m = \{M\},$$
$$\exists_M = \{X \subseteq M : X \neg \varnothing\},$$
$$(\exists_{\geq n})_M = \{X \subseteq M :| \ X \ |\geqq n\},$$
$$(\mathbf{Q}_\alpha)_M = \{X \subseteq M :| \ X \ |\geqq \aleph_\alpha\}, \text{ (the cardinality quantifiers)}$$
$$(\mathbf{Q}_C)_M = \{X \subseteq M :| \ X \ |=| \ M \ |\}, \text{ (the Chang quantifier)}$$
$$(\mathbf{Q}_R)_M = \{X \subseteq M :| \ X \ |>| \ M - X \ |\} \text{ (Rescher's 'plurality quantifier').}$$

All of these satisfy the following condition:

ISOM If f is a bijection from M to M' then $X \in \mathbf{Q}_m \Leftrightarrow f[X] \in \mathbf{Q}_{M'}$.

In fact, Mostowski included *ISOM* as a defining condition on quantifiers, expressing the requirement that 'quantifiers should not allow us to distinguish between element/of M/' [1957, p. 13].

1.4 Generalised Quantifiers

Rescher, introducing the quantifier $\mathbf{Q}_R$, noted that $Q_R x\psi(x)$ expresses

(1) Most things (in the universe) are ψ,

but that the related (and more common)

(2) Most ϕs are ψ

cannot be expressed by means of Q_R [Rescher, 1962]. From our discussion of Frege we recognise (2) a *binary* quantifier, **most**, giving, on each M, a binary relation between subsets of M:

$$\mathbf{most}_M = \{\langle X, Y\rangle \in M^2 :| X \cap Y | > X - Y |\}.$$

To account for this, the construction of formulas must be generalised. This was noted by [Lindström, 1966], who introduced the concept of a *generalised quantifier*, defined below.

(2) can be formalised as

$$\textit{most}\ x, y(\phi(x), \psi(y)).$$

Here the free occurrences of $x(y)$ in $\phi(\psi)$ are bound by the quantifier symbol. In fact, the choice of variables is arbitrary; we can write

$$\textit{most}\ z, x(\phi(x), \psi(x)),$$

or, more simply,

$$\textit{most}\ x(\phi(x), \psi(x)).$$

In this way Mostowskian quantifiers on M are generalised to n-ary relations between subsets of M. A further generalisation is to consider relations between *relations* on M. Here is an example:

$$\mathbf{W}^r_M = \{\langle X, R\rangle : X \subseteq M \& R \subseteq M^2 \& R \text{ wellorders } X\}$$

(The name of this quantifier will be explained later). The statement that (the set) ϕ is wellordered by (the relation) ψ is formalised as

$$W^r x, yz(\phi(x), \psi(y, z))$$

(note that y and z are simultaneously bound in ψ).

Quantifiers are associated with *types* (finite sequences of positive numbers; Mostowskian quantifiers have type $\langle 1\rangle$, **most** has type $\langle 1, 1\rangle$, and $\mathbf{W}^r$ has type $\langle 1, 2\rangle$; the principle should be clear. We are now prepared for the following

DEFINITION 2. A (local) *generalised quantifier of type* $\langle k_1, \ldots, k_n\rangle$ *on* M is an n-ary relation between subsets of $M^{k_1}, \ldots, M^{k_n}$, respectively, i.e. a subset of

$P(M^{k_1}) \times \cdots \times P(M^{k_n})$. A (global) *generalised quantifier of type* $\langle k_1, \ldots, k_n \rangle$ is a function(al) $\mathbf{Q}$ which to each set M assigns a generalised quantifier $\mathbf{Q}_M$ of type $\langle k_1, \ldots, k_n \rangle$ on M. To $\mathbf{Q}$ belongs a quantifier symbol Q (of the same type) with the following rule: If $\phi_1, \ldots, \phi_n$ are formulas and $\bar{x}_1, \ldots, \bar{x}_n$ are strings of distinct variables of length $k_1, \ldots, k_n$, respectively, then $Q\bar{x}_1, \ldots, \bar{x}_n(\phi_1, \ldots, \phi_n)$ is a formula with the truth condition

$$\mathbf{M} \vDash Q\bar{x}_1 \ldots \bar{x}_n(\phi_1, \ldots, \phi_n)[g] \Leftrightarrow \langle \phi_1^{\mathbf{M},g,\bar{x}_1}, \ldots, \phi_n^{\mathbf{N},g,\bar{x}_n} \rangle \in \mathbf{Q}_M.$$

This definition expresses our final version of the relational view of quantifiers, the one we will use in the sequel. It should be clear that, apart from the relativisation to an arbitrary universe M, the notion of a generalised quantifier (or a *Lindström quantifier* as it is sometimes called) is essentially the same as Frege's notion of a second level concept.[11]

Most of the time we will restrict attention to quantifiers of type $\langle 1, 1, \ldots, 1 \rangle$. These are the *monadic generalised quantifiers*; we will usually call them just *quantifiers*. We can then continue to talk about *unary, binary*, etc. quantifiers, when we mean generalised quantifiers of type $\langle 1 \rangle, \langle 1, 1 \rangle$, etc.

Like Mostowski, Lindström included *ISOM* in the definition of generalised quantifiers:

$$\begin{array}{ll} ISOM & \textit{If } f \textit{ is a bijection from } M \textit{ to } M' \textit{ then } \langle R_1, \ldots, R_n \rangle \in \mathbf{Q}_M \\ & \Leftrightarrow \langle f[R_1], \ldots, f[R_n] \rangle \in \mathbf{Q}_{M'}. \end{array}$$

(If R is k-ary, $f[R] = \{\langle f(a_1), \ldots, f(a_k) \rangle : \langle a_1, \ldots, ak \rangle \in R\}$.)

Here are some further examples of generalised quantifiers:

$$\begin{array}{lcl} \mathbf{all}_M & = & \{\langle X, Y \rangle \in M^2 : X \subseteq Y\}, \\ \mathbf{some}_M & = & \{\langle X, Y \rangle \in M^2 : X \cap Y \neq \varnothing\}, \\ \mathbf{I}_M & = & \{\langle X, Y \rangle \in M^2 :\mid X \mid = \mid Y \mid\}, \\ \mathbf{more}_M & = & \{\langle X, Y \rangle \in M^2 :\mid X \mid > \mid Y \mid\}, \\ \mathbf{W}_M & = & \{R \subseteq M^2 : R \text{ wellorders } M\}. \end{array}$$

I is the *Härtig quantifier*, **more** is sometimes called the *Rescher quantifier* (although Rescher only considered the quantifiers $\mathbf{Q}_R$ and **most** above). **W** is the *wellordering quantifier*. The generalised quantifier $\mathbf{W}^r$ given before is the *relativisation* of **W**. This notion is defined as follows.

DEFINITION 3. If $\mathbf{Q}$ is of type $\langle k_1, \ldots, k_n \rangle$, the *relativisation of* $\mathbf{Q}$ is the generalised quantifier $\mathbf{Q}^r$ of type $\langle 1, k_1, \ldots, k_n \rangle$ defined by

$$\langle X, R_1, \ldots, R_n \rangle \in \mathbf{Q}^r_M \Leftrightarrow \langle R_1 \cap X^{k_1}, \ldots, R_n \cap X^{k_n} \rangle \in \mathbf{Q}_X$$

[11] Neither Mostowski nor Lindström seem to have been aware of Frege's concept. there is, however, a tradition within type theory which builds on Frege's work, starting with Church's logic of sense and denotation (cf. [Church, 1951]). More recent works are, e.g. [**?**; Daniels and Freeman, 1978].

(for all $X \subseteq M$ and $R_i \subseteq M^{k_i}$).

Thus for $X \subseteq M$ we can use Q^r to express in M what Q says in X; this will be made precise in 1.6. Note that **all** = $\forall^r$, **some** = $\exists^r$, and **most** = Q^r_R.

1.5 Partially Ordered Prefixes

At this point it is appropriate to mention another generalisation of quantifiers, although not directly related to the relational view. In standard predicate logic each formula can be put in *prenex form*, i.e. with a *linear prefix* $Q_1x_1 \dots Q_nx_n$, where Q_i is either $\forall$ or $\exists$, in front of a quantifier-free formula. Henkin [1961] suggested a generalisation of this to *partially ordered* or *branching* prefixes, e.g. the following

(1)
$$\begin{array}{ll} \forall x & - \exists y \\ & \quad \diagdown \\ & \quad\quad \phi(x,y,z,u) \\ & \quad \diagup \\ \forall z & - \exists u \end{array}$$

The prefix in (1) is called the *Henkin prefix*. The intended meaning of (1) is that for each x there is a y and for each z there is a u such that $\phi(x,y,z,u)$, where y and u are chosen *independently* of one another. To make this precise one uses *Skolem functions*. (1) can then be written

(1′) $\exists f \exists g \forall x \forall z \phi(x, f(x), z, g(z))$.

The method of Skolem functions works for all prefixes with $\forall$ and $\exists$. For example, the first-order

(2) $\forall x \forall z \exists y \exists u \phi(x,y,z,u)$,

(3) $\forall x \exists y \forall z \exists u \phi(x,y,z,u)$

become

(2′) $\exists f \exists g \forall x \forall z \phi(\, f(x,z), z, g(x,z))$.

(3′) $\exists f \exists g \forall x \forall z \phi(x, f(x), z, g(x,z))$.

But the dependencies in (1′) cannot be expressed in ordinary predicate logic; somewhat surprisingly, the Henkin prefix greatly increases the expressive power, as we shall see in 1.6.

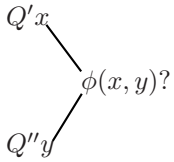
Although branching quantification generalises another feature of ordinary quantification than the one we have been considering here, it can in fact, be subsumed under the relational view of quantifiers. To the Henkin prefix, for example, corresponds the *Henkin quantifier* **H** of type $\langle 4 \rangle$, defined by

$$\begin{aligned} \mathbf{H} = \quad & \{R \subseteq M^4 : \text{ there are functions } f, g \text{ on } M \text{ such that} \\ & \text{for all } a, b \in M, \langle a, f(a), b, g(b) \rangle \in R\}. \end{aligned}$$

The formula (1) is then written, in the notation of 1.4,

(1″) $Hxyzu\phi(x, y, z, u)$.

Observe that branching was only defined for $\forall$ and $\exists$. Can we let other quantifiers branch as well, and consider formulas such as

(4)


It is not immediate what this should mean. Compare the linear

(5) $Q'xQ''y\phi(x, y)$;

this is true in $\mathbf{M}$ iff $X = \{a \in M : \mathbf{M} \vDash Q''y\phi[a, y]\}$ is in $\mathbf{Q}'_M$, and, for each $a \in M, \mathbf{M} \vDash Q''y\phi[a, y]$ iff $Y_a = \{b \in M : \mathbf{M} \vDash \phi[a, b]\}$ is in $\mathbf{Q}''_M$. But the idea with (4) is to evaluate the quantifiers *independently*of each other, and then it is not clear which sets to look for in $\mathbf{Q}'_M$ and $\mathbf{Q}''_M$. Nevertheless, Barwise [1979] shows that for certain Q' and Q'' a reasonable interpretation of (4) can be given, and Westerståhl [1987] extends this to arbitrary Q' and Q''.

Branching quantification is not only of mathematical interest. It can be argued that both the Henkin prefix and the form (4) (for certain non-first-order Q' and Q'') occur essentially in natural languages. Barwise [1979] contains a good presentation of the issues involved here; a brief review will be given in Appendix A.

1.6 Model-Theoretic Logics

The introduction of generalised quantifiers opens up a vast area of logical study. Let *EL* (elementary logic) be standard predicate logic, and, if $\mathbf{Q}^i$ are generalised quantifiers for $i \in I$, let $L(\mathbf{Q}^i)_{i\in I}$ be the logic obtained from *EL* by adding the syntactic and semantic rules for each $\mathbf{Q}^i$ as in Definition 2. The study of such *model-theoretic logics* is sometimes called *abstract model theory*.[12] For a comprehensive survey of this field of mathematical logic the reader is referred to Barwise and Feferman [1985], in particular the chapter [Mundici, 1985]. Below, just a few examples of such logics and their properties will be given.

The expressive power of a logic is most naturally measured by the classes of models its sentences can define. Define $L \leqq L'$ (L' is an *extension* of L) to mean that for each sentence of L there is an equivalent sentence (i.e. one with the same models) of L'. Clearly $\leqq$ is reflexive and transitive, and every logic $L = L(\mathbf{Q}^i)_{i\in I}$ is an extension of *EL*. We write $L \equiv L'$ when $L \leqq L'$ and $L' \leqq L$, and $L < L'$ when $L \leqq L'$ and $L' \not\leqq L$.[13]

[12] There are more general concepts of logic, used in abstract model theory. A comparison of various abstract notions of a logic is given in [Westerståhl, 1976].

[13] This partial order concerns *explicit* power of expression, by *single* sentences. One can also consider *implicit* strength (cf. Appendix B.3), or expressibility by *sets* of sentences.

Since formulas are defined inductively, to prove that $L(\mathbf{Q}^i)_{i\in I} \leqq L'$ it suffices to show that each $\mathbf{Q}^i$ is definable in L'. For example, if $\mathbf{Q}^i$ is of type $\langle 2,1\rangle$ it suffices to show that the sentence

$$Q^i xy, z(P_1 xy, P_2 z)$$

is equivalent to a sentence in L'.

The inductive characterisation of formulas also gives the following result, which explains why *ISOM* is normally assumed for generalised quantifiers in mathematical logic: if each $\mathbf{Q}^i$ satisfies *ISOM*, then truth of sentences in $L(\mathbf{Q}^i)_{i\in I}$ is preserved among isomorphic models. In fact, the inductive proof of this gives slightly more

PROPOSITION 4. *If each* $\mathbf{Q}^i$ *satisfies* ISOM, ϕ *is a formula in* $L(\mathbf{Q}^i)_{i\in I}$, f *an isomorphism from* $\mathbf{M}_1$ *to* $\mathbf{M}_2$, *and* g *an assignment in* M_1, *then*

$$\mathbf{M}_1 \vDash \phi[g] \Leftrightarrow \mathbf{M}_2 \vDash \phi[fg].$$

Here is the relative strength of some of the logics we have considered:

THEOREM 5. $EL < L(\mathbf{Q}_0) < L(\mathbf{I}) < L(\mathbf{more}) < L(\mathbf{H})$.

The easiest part of the proof of this theorem is to show that one logic is an extension of the previous one. That $L(\mathbf{Q}_0) \leqq L(\mathbf{I})$ follows from the equivalence

$$Q_0 xPx \leftrightarrow \exists y(Py \wedge Ix(Px, Px \wedge x \neq y))$$

(P is infinite iff removal of one element does not change its cardinality). That $L(\mathbf{I}) \leqq L(\mathbf{more})$ is obvious, and that $L(\mathbf{more}) \leqq L(\mathbf{H})$ follows by the following trick (due to Ehrenfeucht):

$$\begin{aligned}
\neg \textit{more}\ x(P_1 x, P_2 x) \quad &\leftrightarrow \quad \exists f(f \text{is a 1–1 function from } P_1 \text{ to } P_2) \\
&\leftrightarrow \quad \exists f \forall x \forall z(x = z \leftrightarrow f(x) = f(z) \wedge \\
&\qquad \wedge P_1 x \rightarrow P_2 f(x)) \\
&\leftrightarrow \quad \exists f \exists g \forall x \forall z(x = \leftrightarrow f(x) = g(z) \wedge \\
&\qquad \wedge P_1 x \rightarrow P_2 f(x)) \\
&\leftrightarrow \quad Hxyzu(x = z \leftrightarrow y = u \wedge \\
&\qquad \wedge P_1 x \rightarrow P_2 y).
\end{aligned}$$

To prove that one logic is *not* an extension of another, one can either show directly that some sentence in the first is not equivalent to any sentence in the second, or, more indirectly, use *properties* of the two logics to distinguish them. For example, the following well known properties of *EL* can sometimes be used:

1. *The compactness property: If every finite subset of a set of sentences has a model, the whole set has a model.* Consider the following set of $L(\mathbf{Q}_0)$-sentences:

$$\{\neg Q_0 x(x = x)\} \cup \{\exists_{\geqq n} x(x = x) : n = 1, 2, 3, \ldots\}.$$

This set has no models, but each finite subset has one. So $L(\mathbf{Q}_0)$ (and all its extensions) is not compact. In particular, $L(\mathbf{Q}_0) \not\leqq EL$.

2. *The Tarski property: If a sentence has a denumerable model it has an uncountable model*. Let ϕ be an *EL*-sentence saying that $<$ is a discrete linear ordering with a first element. Then the $L(\mathbf{Q}_0)$-sentence

   (1) $\quad \phi \wedge \forall x \neg Q_0 y(y < x)$

   characterises the natural number ordering $\langle N, < \rangle$ (i.e. $\langle M, R \rangle$ is a model of (1) iff it is isomorphic to $\langle N, < \rangle$). All models of (1) are denumerable, so $L(\mathbf{Q}_0)$ does not have the Tarski property.

3. *The completeness property: The set of valid sentences is recursively enumerable*. Adding to (1) sentences (of *EL*) defining addition and multiplication,and saying that 0 is the least element and $x + 1$ the immediate successor of x, we obtain a sentence θ which characterises the standard model of arithmetic $N = \langle N, <, +, \times, 0, 1 \rangle$. Then, for every $L(\mathbf{Q}_0)$-sentence ψ in this vocabulary,

   $N \vDash \psi \Leftrightarrow \theta \rightarrow \psi$ is valid.

   Thus, since the set of true arithmetical sentences is not recursively enumerable, $L(\mathbf{Q}_0)$ is not complete. This time there is no immediate consequence for extensions of $L(\mathbf{Q}_0)$. For the extensions mentioned in Theorem 5, however, sentences characterising $\boldsymbol{N}$ can be constructed in a similar way, so they are not complete either.

4. *The Löwenheim property: If a sentence has an infinite model it has a denumerable model*. It is not very difficult to show that $L(\mathbf{Q}_0)$ in fact has the Löwenheim property. But $L(\mathbf{I})$ (and its extensions) does not: we can write down a sentence of $L(\mathbf{I})$ saying that $<$ is a dense linear ordering without endpoints, and that there is an element which does not have as many predecessors s it has successors. In a model, the set of predecessors and the set of successors of this element are infinite and of different cardinalities, so the model must be uncountable. It follows, in particular, that $L(\mathbf{I}) \not\leqq L(\mathbf{Q}_0)$.

In the proof of Theorem 5, it only remains to show that $L(\mathbf{I})$ is not an extension of $L(\mathbf{more})$, and that $L(\mathbf{more})$ is not an extension of $L(\mathbf{H})$. A convenient way to prove the former will be given in 1.7. A proof of the latter can be found in [Cowles, 1981].

Recall the definition of relativised quantifiers in Section 1.4.2. We say that $L = L(\mathbf{Q}^i)_{i \in I}$ *relativises*, if

$$L^r = L((\mathbf{Q}^i)^r)_{i \in I} \leqq L,$$

i.e. if the relativisation of each $\mathbf{Q}^i$ is definable in L. $EL, L(\mathbf{Q}_\alpha), L(\mathbf{I}), L(\mathbf{most})$, $L(\mathbf{more})$ and $L(\mathbf{H})$ all relativise. For example,

$$\begin{aligned}&\forall^r x(Px, P_1x) \leftrightarrow \forall x(Px \rightarrow P_1x),\\ &\textit{most}^r\, x(Px, P_1x, P_2x) \leftrightarrow \textit{most}\, x(Px \wedge P_1x, P_2x),\\ &H^r v, xyzu(Pv, P_1xyzu) \leftrightarrow Hxyzu((Px \wedge Pz) \rightarrow\\ &\quad (Py \wedge Pu \wedge P_1xyzu)).\end{aligned}$$

$L(\mathbf{Q}_R), L(\mathbf{Q}_C)$ and $L(\mathbf{W})$, on the other hand, do not relativise (cf. Section 1.7).

As the above equivalences show, relativised quantifier symbols are used to make relativised statements. This extends to all L-sentences. Define, for each L-formula ϕ and each unary predicate symbol P, the *relativised formula*

$$\phi^{(P)}$$

In L^r inductively by letting $\phi^{(P)} = \phi$ if ϕ is atomic, $(\neg\psi)^{(P)} = \neg\psi^{(P)}, (\psi \wedge \theta)^{(P)} = \psi^{(P)} \wedge \theta^{(P)}$, and, when ϕ is quantified, beginning with Q^i of type $\langle 2, 1\rangle$, say,

$$Q^i xy, (\psi, \theta)^{(P)} = (Q^i)^r v, xy, z(Pv, \psi^{(P)}, \theta^{(P)}).$$

$\phi^{(P)}$ expresses exactly what ϕ says about the universe restricted to (the denotation of) P. We can formulate this precisely as follows. Call a subset X of the universe of the model $\mathbf{M}$ *universe-like* if $X \neq \varnothing$, the denotations of all individual constants in the vocabulary for $\mathbf{M}$ are in X, and X is closed under the denotations of all function symbols in the vocabulary. In that case, let $\mathbf{M} \mid X$ be the model with universe X, and all the relations etc; in $\mathbf{M}$ restricted to X. Then it can be shown by induction that if X is universe-like and ϕ is an L sentence,

(*REL*) $\quad (\mathbf{M}, X) \vDash \phi^{(P)} \Leftrightarrow \mathbf{M} \mid X \vDash \phi$

(here we assume that P does not occur in ϕ and that it denotes X in $(\boldsymbol{M}, X)$).

If L relativises, all this can be done in L, since $\phi^{(P)}$ is then clearly equivalent to an L-sentence.

So far we have only discussed particular logics and their properties. The most exciting part of abstract model theory, however, concerns results relating various properties of logics to each other, and results *characterising* certain logics in terms of their properties. Most famous of these characterisations is still *Lindström's theorem* [1969], which characterises *EL* in terms of the four properties mentioned above (for proofs, cf. [Flum, 1985], van Benthem and Doets or Hodges (both this Handbook series).

THEOREM 6. *If L is compact and has the Löwenheim property, then $L \equiv EL$. Also, of L relativises, then* (a) *if L is complete and has the Löwenheim property then $L \equiv EL$;* (b) *if L has the L′owenheim and Tarski property then $L \equiv EL$.*

1.7 The Strength of Monadic Quantifiers

In general, it may be quite difficult to determine whether $L \leqq L'$ or not, where L and L' are logics with generalised quantifiers. In the case of *monadic* quantifiers, however, things become much easier. Since this case is what we shall mainly be dealing with, I will devote the present subsection to developing some machinery for comparing the expressive power of logics with monadic quantifiers. The machinery will be applied in particular to the quantifiers **more** and **most**. I use these quantifiers later to illustrate some important points concerning natural language quantification, and it will then be instructive to have established their logical properties.

This subsection is a bit more technical than the previous ones; I have written out proofs of results that are new or not easily found in the literature (cf. the bibliographical note at the end). The reader can skip or glance through it now, and return to it for a definition or a result that is used later.

From now on, when $\mathbf{Q}$ is an m-ary monadic quantifier, we will write simply

$$\mathbf{Q}_M X_1 \dots X_m,$$

instead of $\langle X_1, \dots, X_m \rangle \in \mathbf{Q}_M$. Thus,

$$\begin{aligned} &\mathbf{all}_M AB \Leftrightarrow A \subseteq B, \\ &\mathbf{most}_M AB \Leftrightarrow |A \cap B| > |A - B|, \\ &\mathbf{more}_M AB \Leftrightarrow |A| > |B|, \end{aligned}$$

etc.

Let $\mathbf{M} = \langle M, A_0, \dots, A_{k-1} \rangle$ be a *K-ary monadic structure* (i.e. the A_i are subsets of M, and the vocabulary consists of k unary predicate symbols). The following terminology will be used her an in later sections. If $X \subseteq M$, let $X^0 = X$ and $X^1 = M - X$. If s is a function from $\{0, \dots, k-1\}$ to $\{0, 1\}$, i.e. if $s \in 2^k$, let

$$P_s^M = A_0^{s(0)} \cap \dots \cap A_{k-1}^{s(k-1)}.$$

$\{P_s^M\}_{s \in 2^k}$ is a *partition* of M, and, up to isomorphism, the number of elements in these partition sets is all there is to say about $\mathbf{M}$. In other words, if $|P_s^{\mathbf{M}}| = |P_s^{\mathbf{M}'}|$ for all $s \in 2^k$, then $\mathbf{M}$ and $\mathbf{M}'$ are isomorphic. Finally, let

$$U_i^{\mathbf{M}},$$

for $1 \leqq i \leqq 2^{2^k}$, be all possible *unions* of the partition sets (including $\varnothing$), in some fixed order.

If L is a logic, $\mathbf{M}$ a structure (not necessarily monadic), $X \subseteq \mathbf{M}$, and $a_1, \dots, a_n \in M$, X is said to be *L-definable in $\mathbf{M}$ with parameters $a_1, \dots, a_n$*, if there is an L-formula ϕ in the vocabulary of $\mathbf{M}$ such that

$$a \in X \Leftrightarrow \mathbf{M} \vDash \phi[a, a_1, \dots, a_n].$$

The following is an almost immediate consequence of this definition and Proposition 4:

LEMMA 7. *If L satisfies ISOM, X is L-definable in $\mathbf{M}$ with parameters $a_1, \ldots, a_n$, and f is an automorphism on $\mathbf{M}$ (i.e. an isomorphism from $\mathbf{M}$ to $\mathbf{M}$) with $f(a_i) = a_i$, then $f[X] = X$.*

If A, B are sets, $A \oplus B$, the *symmetric difference* between A and B, is $(A - B) \cup (B - A)$. We say that B is an X-*variant* of A, if $A \oplus B \subseteq X$.

LEMMA 8. *Suppose that L satisfies ISOM and that $\mathbf{M}$ is a monadic structure. Then the L-definable sets in $\mathbf{M}$ with parameters $a_1, \ldots, a_n$ are precisely the $\{a_1, \ldots, a_n\}$-variants of the unions $U_i^{\mathbf{M}}$.*

Proof. Clearly all these sets are also definable. Now suppose X is L-definable in $\mathbf{M}$ from $a_1, a \ldots, a_n$. Then so is $X' = X = \{a_1, \ldots, a_n\}$. It suffices to show that X' has the desired form. Let $s_1, \ldots, s_p$ be those $s \in 2^k$ for which $X' \cap P_s^m \neq \varnothing$. Thus,

$$X' \subseteq P_{s_1}^M \cup \ldots \cup P_{s_p}^M.$$

Suppose X' is *not* and $\{a_1, \ldots, a_n\}$-variant of $P_{s_1}^M \cup \ldots \cup P_{s_P}^M$. Then, for some i, there is $a \in P_{s_i}^M - X'$ such that $a \neq a_1, \ldots, a_n$. But, by the construction, there is $b \in P_{s_i}^M \cap X$; such that $b \neq a_1, \ldots, a_n$. let $f(a) = b, f(b) = a$, and $f(x) = x$ when $x \neq a, b$. Then f is an automorphism on $\mathbf{M}$ leaving $a_1, \ldots, a_n$ fixed, so $f[X'] = X'$, by Lemma 7. But this contradicts the fact that $a \in f[X'] - X'$. ■

Now we restrict attention to logics with monadic quantifiers satisfying ISOM. For simplicity, assume that $L = L(\mathbf{Q})$, where $\mathbf{Q}$ is *binary*; the results below extend immediately to logics $L(\mathbf{Q}^o \mapsto \mathbf{Q}^i)_{i \in I}$, with monadic $\mathbf{Q}^i$.

The *quantifier rank* of L-formulas is defined inductively as follows:

$$\begin{aligned}
&qr(\phi = 0, \text{ if } \phi \text{ is atomic,}\\
&qr(\neg\phi) = qr(\phi)\\
&qr(\phi \wedge \psi) = \mathit{max}(qr(\phi), qr(\psi)),\\
&qr(\exists x\phi) = qr(\phi + 1)\\
&qr(Qx(\phi, \psi)) = \mathit{max}(qr(\phi), qr(\psi)) + 1.
\end{aligned}$$

we write

$$\mathbf{M} \equiv_{n,Q} \mathbf{M}'$$

to mean the same $L(\mathbf{Q})$-sentences of quantifier rank at most n are true in $\mathbf{M}$ and $\mathbf{M}'$. $\mathbf{M} = \equiv_Q \mathbf{M}'$ ($\mathbf{M}$ and $\mathbf{M}'$ are $L(\mathbf{Q})$-*equivalent*) if, for all n, $\mathbf{M} \equiv_{n,Q} \mathbf{M}'$. Our main tool will be an equivalent but more workable formulation of the relation $\equiv_{n,Q}$. This is accomplished in the next definition. If $a_1, \ldots, a_n \in M$ and $b_1, \ldots, b_n \in M'$ we write $(a_1, \ldots, a_n) \simeq_p (b_1, \ldots, b_n)$ to mean that $\{\langle a_i, b_i\rangle :$

$1 \leqq i \leqq n\}$ is a *partial isomorphism* from $\mathbf{M}$ to $\mathbf{M}'$ (i.e. $a_i = a_j$ iff $b_i = b_j$, and $a_i \in A_m$ iff $b_i \in A'_m$).

In what follows, $\mathbf{M}$ and $\mathbf{M}'$ are k-ary monadic structures.

DEFINITION 9.

(a) $X \approx_n Y$ iff either $|X| = |Y| < n$ or $|X|, |Y| \geqq n$.

(b) $\mathbf{M} \approx_n \mathbf{M}'$iff $P_s^{\mathbf{M}} \approx_n P_s^{\mathbf{M}'}$ for all $s \in 2^k$

(c) $\mathbf{M} \approx_{n,Q} M'$ iff

(i) $\mathbf{M} \approx_n \mathbf{M}'$

(ii) If $(a_1, \ldots, a_{n-1}) \simeq_p (b_1, \ldots, b_{n-1}), X_i, X_j$ are $\{a_1, \ldots, a_{n-1}\}$-variants of $U_i^{\mathbf{M}}, U_j^{\mathbf{M}}$, an Y_i, Y_j the *corresponding* $\{b_1, \ldots, b_{n-1}\}$-variants of $U_i^{\mathbf{M}'}, U_j^{\mathbf{M}'}$, then

$$\mathbf{Q}_M X_i X_j \Leftrightarrow \mathbf{Q}_{M'} Y_i Y_j.$$

THEOREM 10. $\mathbf{M} \equiv_{n\mathbf{Q}} \mathbf{M}' \Leftrightarrow \mathbf{M} \approx_{n,\mathbf{Q}} \mathbf{M}'$.

Proof. $\Rightarrow$: It is clear that (i) holds. As for (ii), let $\psi_i(y, x_1, \ldots, x_{n-1}), \psi_j(y, x_1, \ldots, x_{n-1})$ be formulas which L-define X_i, X_j in $\mathbf{M}$ with parameters $a_1, \ldots, a_{n-1}$). Each a_p belongs to exactly one $P_{s_p}^{\mathbf{M}}$; let this set be defined by $\theta_p(x)$. If $\mathbf{Q}_{\mathbf{M}} X_i, X_j$, then

$$\mathbf{M} \vDash \exists x_1, \ldots, x_{n-1}(\theta_1(x_1) \wedge \ldots \wedge \theta_{n-1}(x_{n-1}) \wedge \\ \wedge Qy(\psi_i(y, x_1, \ldots, x_{n-1}), \psi_j(y, x_1, \ldots, x_{n-1}))).$$

This sentence has quantifier rank n. Thus, it is also true in $\mathbf{M}'$, whence there are $b'_1, \ldots, b'_{n-1} \in M'$ such that $b'_p \in P_{s'_p}^{\mathbf{M}'}$ and

$$\mathbf{M}' \vDash Qy(\psi_i, \psi_j)[b', \ldots, b'_{n-1}].$$

Let f map b'_p on b_p and leave everything else in M' as it is. It follows that f is an automorphism on $\mathbf{M}'$, so

$$\mathbf{M}' \vDash Qy(\psi_i, \psi_j)[b_1, \ldots, b_{n-1}].$$

but this means that $\mathbf{Q}_{M'} Y_i, Y_j$. The converse is similar.

$\Leftarrow$: We prove by (downward) induction over $p \leqq n$ that

($*$) If $(a_1, \ldots, a_p) \simeq_p (b_1, \ldots, b_p)$ and $qr(\phi) \leqq n-p$, then $\mathbf{M} \vDash \phi[a_i, \ldots, a_p] \Leftrightarrow \mathbf{M}' \vDash \phi[b_1, \ldots, b_p]$.

The case $p = 0$ gives the result. $(*)$ is clear for $p = n$. So suppose $(*)$ holds for $p, (a_1, \ldots, a_{p-1}) \simeq_p (b_1, \ldots, b_{p-1})$ and $qr(\phi) = n - p + 1$. We may suppose that ϕ begins with a quantifier symbol. If this symbol is $\exists$, the result follows easily from the induction hypothesis and the fact that $\mathbf{M} \approx_n \mathbf{M}'$. So suppose ϕ is $Qx(\psi_1, \psi_2)$. Let $\psi_i^{\mathbf{M}} = \{a \in M : \mathbf{M} \vDash \psi_i[a, a_1, \ldots, a_{p-1}]\}, i = 1, 2$, and similarly for $\psi_i^{\mathbf{M}'}$. By Lemma 8, each ψ_i^M is an $\{a_1, \ldots, a_{p-1}\}$-variant of some union $U_{j_i}^{\mathbf{M}}$ of partition sets.

CLAIM: $\psi_i^{\mathbf{M}'}$ is the corresponding $\{b_1, \ldots, b_{p-1}\}$-variant of $U_{j_i}^{\mathbf{M}'}$.

The result follows immediately from the chain and (ii) above. The proof of the claim is straightforward, using the induction hypothesis together with the fact that $\mathbf{M} \approx_n \mathbf{M}'$. ■

As noted, the theorem extends to logics with several monadic quantifiers (satisfying ISOM). We use this in the next corollary. A k-ary quantifier $\mathbf{Q}$ is said to be *closed under* $\approx_{n,\mathbf{Q}^1 \ldots \mathbf{Q}^m}$ if $\mathbf{Q}_M A_0 \ldots A_{k-1}$ and $\langle M, A_0, \ldots, A_{k-1}\rangle \approx_{n,\mathbf{Q}^1,\ldots,\mathbf{Q}^m} \langle M', A_0', \ldots, A_{k-1}'\rangle$ implies $\mathbf{Q}_{\mathbf{M}'} A_0', \ldots A_{k-1}'$.

COROLLARY 11. *A monadic quantifier* $\mathbf{Q}$ *is definable in* $L(\mathbf{Q}^1, \ldots, \mathbf{Q}^m)$ *if and only if, for some natural number* n, $\mathbf{Q}$ *is closed under* $\approx_{n,\mathbf{Q}^1 \ldots \mathbf{Q}^m}$.

Proof.[outline] If $\mathbf{Q}$is defined by a sentence ϕ in $L(\mathbf{Q}^1, \ldots, \mathbf{Q}^m)$, i.e. if

$$\mathbf{Q}_M A_0, \ldots, A_{k-1} \Leftrightarrow \langle M, A_0, \ldots, A_{k-1}\rangle \vDash \phi,$$

just let n be the quantifier rank of ϕ and use the theorem. Conversely, note that, with a fixed finite vocabulary there are, up to logical equivalence, only finitely many $L(\mathbf{Q}^1, \ldots, \mathbf{Q}^m)$-sentences of quantifier rank at most n. Now take the disjunction of all such sentences which are 1-complete n-descriptions of the models $\langle M, A_0, \ldots, A_{k-1}\rangle$ for which $\mathbf{Q}_M A_0, \ldots, A_{k-1}$; this disjunction defines $\mathbf{Q}$. ■

We will now apply these results to some particular monadic quantifiers. First, note the following special cases of Theorem 10:

1. $\mathbf{M} \equiv_n \mathbf{M}' \Leftrightarrow \mathbf{M} \approx_n \mathbf{M}'$,
2. If $\mathbf{Q}$ is first-order definable, then $\mathbf{M} \equiv_{n,\mathbf{Q}} \mathbf{M}' \Leftrightarrow \mathbf{M} \approx_n \mathbf{M}'$.

Using this, one easily shows that quantifiers such as $\mathbf{Q}_\alpha, \mathbf{Q}_C, \mathbf{Q}_R$ are *not* first-order definable. Next, note that an $\{a_1, \ldots, a_{n-1}\}$-variant of $U_i^{\mathbf{M}}$ has cardinality $\geqq \aleph_\alpha$ iff $U_i^{\mathbf{M}}$ has cardinality $\geqq \aleph_\alpha$ iff one of the partition sets in U_i^M has cardinality $\geqq \aleph_\alpha$. Thus, when $\mathbf{Q} = \mathbf{Q}_\alpha$, we need only consider the partition sets (not variants of unions of them) in Definition 9(c). This makes it easy to show,for example, that if $\alpha \neq \beta$, $L(\mathbf{Q}_\alpha)$ and $L(\mathbf{Q}_\beta)$ have *incomparable* expressive power.

3. $L(\mathbf{Q}_R \nleqq L(\mathbf{I})$.

Proof. By the theorem, it suffices to find, for each n structures $\langle M, A\rangle$ and$\langle M', A'\rangle$ such that $\langle M, A\rangle \equiv_{n,I} \langle M', A'\rangle, (\mathbf{Q}_R)_M A$, and $\neg(\mathbf{Q}_R)_{M'} A'$. But this is easy. For example, let $|A| = 4n, |M - A| = 2n, |A'| = 2n, |M' - A;| = -4n$. There are just four unions of partition sets to consider in each structure, and it is easy to verify that the conditions in Definition 9(c) are satisfied. ■

4. "$|A|$ is even" is not expressible in L(**more**).

Proof. For each n, choose $M, M', A \subseteq M, A' \subseteq M'$ such that $|A| = 4n, |M - A| = |M' - A'| = n, |A'| = 4n + 1$. Then $\langle M, A\rangle \approx_{n,\mathbf{more}} \langle M', a'\rangle$, so $\langle M, A\rangle \equiv_{n,\mathbf{more}} \langle M', a'\rangle$ by the theorem, but $|A|$ is even and $|A'|$ is odd. ■

The following result is from Barwise and Cooper [1981]:

5. $L(\mathbf{most}) \not\leqslant L(\mathbf{Q}_R)$, i.e. $L(\mathbf{Q}_R)$ does not relativise.

Proof. Given n, choose $\langle M, A_0, A_1\rangle, \langle M', A'_0, A'_1\rangle$ such that $A_0 \cap A_1 = \varnothing, A'_0 \cap A'_1 = \varnothing, |A_0| = |A_1| = n, |M| = 6n, |A'_0| = n, |A'_1| = n + 1, |M'| = 6n + 2$. So $\varnothing, A_0, A_1, A_0 \cup A_1$ all have cardinalities less than their complements, and this continues to hold if $n - 1$ elements are 'moved around' in the model. The same holds for $\mathbf{M}'$, and it is then easy to see that $M \equiv_{n,Q_R} M'$. However, $\neg$**most**$_M, A_0 \cup A_1 A_1$ and **most**$_{M'} A'_0 \cup A'_1 A'_1$. ■

Similarly, we can prove that $\mathbf{Q}_C$ does not relativise. Note that only finite structures have been used so far. The next and final application involves infinite structures.

6. $L(\mathbf{Q}_0) \not\leqslant L(\mathbf{most})$.

Proof. This time, choose $\langle M, A\rangle, \langle M', A'\rangle$ such that $|M - A| = |M' - A'| = n, |A| = \aleph_0$, and $|A'| = 3n$. Again, it is not hard to see that $\langle M, A\rangle \approx_{n,\mathbf{most}} \langle M', A'\rangle$ (especially if we use the characterisation of $\approx_{n,\mathbf{most}}$ given in Theorem 12 below), but A is infinite and A' is finite. ■

Finally, we shall consider more closely the relative expressive power of **most** and **more**. Note first that the four properties of logics mentioned in Section 1.6 do not enable us to distinguish between these two quantifiers: we saw that $L(\mathbf{more})$ does not have any of these properties, and similar arguments establish that neither does $L(\mathbf{most})$. For example, if we replace the second conjunct in the sentence (1) in Section 1.6 by a sentence saying that, for each x (except the first) there is a *greatest* element $y < x$ with the property that most of the x-predecessors are not predecessors of y, then we again obtain a characterisation of the natural number ordering.

The next result characterises the relations $\equiv_{n,\mathbf{Q}}$ and $\equiv_{\mathbf{Q}}$ for monadic structures, when $\mathbf{Q}$ is **most** or **more**.

THEOREM 12.

(a) $\mathbf{M} \equiv_{n,\mathbf{more}} \mathbf{M}'$ *iff, whenever* $(a_1, \ldots, a_{n-1}) \simeq_p (b_1, \ldots, b_{n-1}), X_i, X_j$ *are* $\{a_1, \ldots, a_{n-1}\}$*-variants of* $U_i^{\mathbf{M}}, U_j^{\mathbf{M}}$ *and* Y_i, Y_j *the corresponding* $\{b_1, \ldots, b_{n-1}\}$*-variants of* $U_i^{\mathbf{M}'}, U_j^{\mathbf{M}'}$*, we have* $|X_i| > |X_j| \Leftrightarrow |Y_i| > |Y_j|$.

(b) For $\equiv_{n,\mathbf{most}}$ *we have the same condition, except that* $X_i, X_j (Y_i, Y_j)$ *are required to be disjoint.*

(c) $\mathbf{M} =_{\mathbf{more}} \mathbf{M}'$ *iff* $\mathbf{M} \equiv_{\mathbf{most}} \mathbf{M}'$ *iff* $\mathbf{M} \equiv_{\aleph_0} \mathbf{M}'$ *and, for all* $s, t \in 2^k$, $|P_s^{\mathbf{M}}| > |P_t^{\mathbf{M}}|$ *od* $\Leftrightarrow |P_s^{\mathbf{M}'}| > |P_t^{\mathbf{M}'}|$.

Proof.

(a) This is Theorem 10, except that we must show that the condition on the right hand side of the equivalence ((ii) in Definition 9 (c)) implies that $\mathbf{M} \approx_n \mathbf{M}'$. So suppose first $|P_s^{\mathbf{M}}| < n$. Suppose also that $|P_s^{\mathbf{M}'}| \neq |P_s^{\mathbf{M}}|$, say $|P_s^{\mathbf{M}'}| < |P_s^{\mathbf{M}}|$ (the other case is similar). If $P_s^{\mathbf{M}'} = \{b_1, \ldots, b_r\}$, choose $a_1, \ldots, a_r \in P_s^{\mathbf{M}}$ and let $Y_i = \varnothing = P_s^{\mathbf{M}'} - \{b_1, \ldots, b_r\}$ and $Y_j = \varnothing$. It follows from the condition that $X_i = P_s^{\mathbf{M}} - \{a_1, \ldots, a_r\}$ is empty, contradicting our assumption. The case when $|P_s^{\mathbf{M}}| \geqq n$ is similar.

(b) From left to right, note that **most** allows us to compare the cardinalities of *disjoint* sets $X, Y \subseteq M$: then $|X| < |Y|$ iff $\mathbf{most}_M X \cup Y X$. In the other direction, observe first that the argument in (a) above goes through under the disjointness requirement. Moreover, the proof of Theorem 10 ($\Leftarrow$:) also goes through under this requirement, since the formula *most* $x(\psi_1, \psi_2)$ only 'compares' disjoint sets.

(c) Clearly $\mathbf{M} \equiv_{\mathbf{more}} \mathbf{M}'$ implies $\mathbf{M} \equiv_{\mathbf{most}} \mathbf{M}'$, which in turn implies the rightmost condition in (c). Now suppose that condition holds; we must show that, for all $n, \mathbf{M} \approx_{n,\mathbf{more}} \mathbf{M}'$. So take n, and suppose $a_1, \ldots, a_{n-1}, b_1, \ldots, b_{n-1}, X_i, X_j, Y_i, Y_j$ are as in (a) above. We assume $|X_i| > |X_j|$ and show that , in this case, $|Y_i| > |Y_j|$; the other direction is similar.

Case 1: X_i and X_j are both finite. Then the partition sets in $U_i^{\mathbf{M}}$ are finite and thus have the same cardinality as the corresponding partition sets in $U_i^{\mathbf{M}'}$, since $\mathbf{M} \approx_n \mathbf{M}'$ for all n. X_i differs from $U_i^{\mathbf{M}'}$ only by certain of the $a_1, \ldots, a_{n-1}$, and Y_i differs in the same way from $U_i^{\mathbf{M}'}$. Therefore, $|X_i| = |Y_i|$ and $|X_j| = |Y_j|$, and the conclusion follows.

Case 2: X_i and X_j are both infinite. Then $|X_i|$ is the *max* of the cardinalities of the partition sets making up $U_i^{\mathbf{M}}$; say, $|X_i| = |P_s^{\mathbf{M}}|$, and similarly $|X_j| = |P_t^{\mathbf{M}}|$. It then follows from the condition in (c) that $|Y_i| = |P_s^{\mathbf{M}'}|$ and $|Y_j| = P_t^{\mathbf{M}'}|$. Since $|P_s^{\mathbf{M}}| > |P_t^{\mathbf{M}}|$ we have, again by the condition, $|P_s^{\mathbf{M}'}| > |P_t^{\mathbf{M}'}|$.

Case 3: X_i is infinite and X_j is finite. Arguing as in Cases 1 ad 2, we see that Y_i is infinite and Y_j is finite. ■

Thus, the relations $\equiv_{\mathbf{most}}$ and $\equiv_{\mathbf{more}}$ coincide on monadic structures (but *not* the relations $\equiv_{n,\mathbf{most}}$ and $\equiv_{n,\mathbf{more}}$). Nevertheless, $L(\mathbf{more})$ is more expressive than $L(\mathbf{most})$, even if we restrict attention to monadic structures, as the next result will show. Another instance of the same phenomenon is given by the fact that

$$\mathbf{M} \equiv_{Q_0} \mathbf{M}' \Leftrightarrow \mathbf{M} \equiv \mathbf{M}'$$

(this is an easy consequence of Theorem 10), but $EL < L(\mathbf{Q}_0)$ (even on monadic structures).

The following theorem holds in general, but it is also true if only monadic structures are considered.

THEOREM 13.

(a) $L(\mathbf{most}) < L(\mathbf{more})$.

(b) $L(\mathbf{most}) \equiv L(\mathbf{more})$ on finite structures.

(c) $L(\mathbf{more}) \equiv L(\mathbf{most}, \mathbf{Q}_0)$.

Proof.

(a) Clearly $L(\mathbf{most}) \leqq L(\mathbf{more})$. That $L(\mathbf{more}) \not\leqq L(\mathbf{most})$ follows from (6) and Theorem 5.

(b) This follows from the fact that, if $A \cap B$ is finite, then $\mathbf{more}_M AB \Leftrightarrow |A| > |B| \Leftrightarrow |A - B| > |B - A| \Leftrightarrow \mathbf{most}_M A \oplus BA$.

(c) We must show that $L(\mathbf{more}) \leqq L(\mathbf{most}, \mathbf{Q}_0)$. Take any M and $A, B \subseteq M$. If $A \cap B$ is finite, $\mathbf{more}_M AB$ is expressed as in (B). If $A \cap B$ is infinite, then $|A| = max(|A - B|, |A \cap B|)$ and $|B| = max)|B - A|, |A \cap B|)$. It follows that

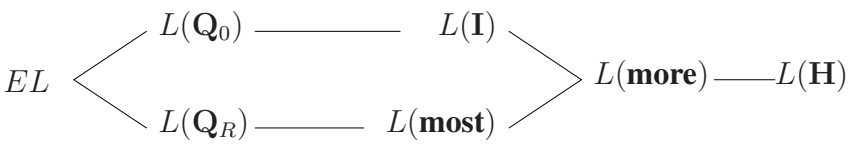
$$|A| > |B| \Leftrightarrow |A - B| > |B - A| \& |AB| > |A \cap B|,$$

and the right hand side of this is again expressible with *most* (since only disjoint sets are compared). Moreover, $\mathbf{Q}_0$ allows us to distinguish the two cases, in one sentence of $L(bfmost, \mathbf{Q}_0)$.

■

This theorem tells us that the difference between $L(\mathbf{more})$ and $L(\mathbf{most})$ is *precisely* that the former, but not the latter, can distinguish between infinite and finite sets.

The results of this section allow us to extend Theorem 5 to the following picture:



Here each logic is strictly stronger than its immediate predecessor(s), and logics not on the same branch are incomparable.

REMARK 14. The only thing in the figure above not proved with the simple methods used here is the fact that $L(\mathbf{H})$ is strictly stronger than $L(\mathbf{more})$. However, if we consider the logic L^{po}, where not only the Henkin prefix but *all* partially ordered prefixes with $\forall$ and $\exists$ are allowed, then it follows from (4) that $L^{po} \not\leqq L(\mathbf{more})$. For,

$$|A| \text{ is even } \Leftrightarrow \exists X \subseteq A(|X| = |A - X|),$$

which can be expressed as a Σ^1_1 sentence, and is shown in [Enderton, 1970] and [Walkoe, 1970] that all such sentences are expressible in L^{po}.

Is '$|A|$ is even |' expressible in $L(*\mathbf{H})$? More generally, is $L(\mathbf{H})$ strictly stronger than $L(\mathbf{more})$ if we restrict attention to monadic structures/ I don't know the answer to these questions, but it may be noted that it follows from Theorem 12 and a result in [Lachlan and Krynicki, 1979] that $\equiv_{\mathbf{more}}$ and $\equiv_{\mathbf{H}}$ coincide for monadic structures.

Bibliographical remark: The theorems in this section have not, to my knowledge, appeared in the literature, although no doubt they belong to the folklore in some circles. Most of the applications to particular logics are known, but it should be noted that the methods used here are much more elementary than the ones that have been used in the literature the proof of (5) in [Barwise and Cooper, 1981] is an exception). For example, it is proved in [Hauschild, 1981] and [Weese, 1981] that $L(\mathbf{more})$ is strictly stronger than $L(\mathbf{I})$ by establishing that these logics have different properties w.r.t. the decidability of certain theories formulated in them. The same result follows from the simple observation 93); in a sense, (3) gives more, since it concerns monadic structures, whereas the theories just mentioned use non-monadic languages.

2 NATURAL LANGUAGE QUANTIFIERS

A main objective of Montague's PTQ [Montague, 1974] was to show that intensional phenomena, such as quantification into intensional contexts, could be handled rigorously with model-theoretic methods. But even if one completely disregards the intensional aspects of PTQ, its approach to quantification was novel. Although it had no category 'quantifier' or 'determiner', a general pattern is discernible from its treatment of the three quantifier expressions (*every, a*, and *the*)

it in fact did account for. The basic idea is that *quantifier expressions occur as determiners in noun phrases*. By the close correspondence between syntax and semantics in Montague Grammar, this also determines the interpretation of such expressions.

In this section, I will describe this idea in somewhat more detail, and its later development in [Barwise and Cooper, 1981] and [Keenan and Stavi, 1986], within the generalised quantifier framework of Section 1.

2.1 Determiners

Suppose that the expressions of the categories *common noun* (N) and *noun phrase* (NP) have somehow been (roughly) identified.[14] Since we are disregarding intensions,the semantic types of these expressions are such that Ns are interpreted, in a model $\mathbf{M} = \langle M, \| \, \| \rangle$ with universe M and interpretation function $\| \, \|$, as *subsets* of M an NPs as *sets of subsets* of M. Here are three examples from PTQ:

$$\|\textit{every man}\| = \{X \subseteq M : \|\textit{man}\| \subseteq X\},$$
$$\|\textit{a man}| = \{X \subseteq M : \|\textit{man}\| \cap X \neq \varnothing\},$$
$$\|\textit{the man}\| = \{X \subseteq M : |\|\textit{man}\|| = 1 \& \|\textit{man}\| \subseteq X\}.$$

Many NPs, like the above ones, are naturally regarded as the result of applying a syntactic *operator* to Ns. We introduce the syntactic category *determiner* (*DET*) for this sort of operator:

(*DET*) *DETs form NPs from Ns.*

This is a rough criterion, but, in a Montagovian framework, it is enough to fix the syntax and semantics of determiners. In particular, *DET*s are interpreted as *functions* from N denotations to NP denotations. For example,

$$\|\textit{every}\|(A) = \{X \subseteq M : A \subseteq X\},$$
$$\|\textit{a}\|(A) = \{X \subseteq M : A \cap X \neq \varnothing\},$$
$$\|\textit{the}\|(A) = \{X \subseteq M : |A| = 1 \& A \subseteq X\}.$$

Another thing, of course, is to apply the criterion to identify simplex and complex English *DET*s; we will return to this in Section 2.4.

2.1.1 Three apparent problems

As noted, the basic idea of the present Montague-style treatment of quantification is this:

(*Q*) *Quantifier expressions are DETs.*

[14] We don't need to assume that proper *definitions* of these categories exist, only that there is agreement about them in a large number of cases.

This may not yet seem very exciting, but note at least that it differs, syntactically as well as semantically, from the standard predicate logic treatment of quantification. The import of (Q) will become clear as we go along. For the moment, however, let us look at a few apparent *counter-instances* to (Q) that come to mind.

I. In sentences like

(1) All cheered,

(2) Some like it hot,

(3) Few were there to meet him,

the words *all, some, few* are not applied to arguments of category N. Isn't the standard predicate logic analysis more plausible here? No, it is very natural to assume that the *DET*s have 'dummy' arguments in these sentences (what context-given interpretations); in this case (Q) still holds (cf. 2.4.5).

II. Words like *something, everything, nothing, nobody*, etc. look like quantifier expressions but are certainly not *DET*s. We have two options here. The first is to regard them as simplex *NP*s, denoting quantifiers of type $\langle 1 \rangle$ in the sense of 1.4. They would then correspond (roughly) to the standard logical $\forall$ and $\exists$. The other option, which we will take here, is to regard them as *complex: something = some(thing), nothing = no(thing)*, etc.; i.e. obtained by applying a *DET* to a (perhaps logical) N like *thing*. In this way, (Q) can be maintained.

III. In 1.4 we defined the binary quantifier **more**. The word *more*, however, is not a *DET* by our criterion; compare

(4) Some boys run,

(5) Most boys run,

(6) *More boys run.[15]

Still, *more* does occur in quantified sentences, for example,

(7) There are more girls than boys,

which in generalised quantifier notation becomes

(8) *more* $x(\mathit{girl}(x), \mathit{boy}(x))$.

[15]Even if there are contexts where (6) might be uttered, it is unreasonable to interpret *more* as an independent *DET*: the standard of comparison is missing, and has to be supplied to get at the meaning. So *more* in (6) would then stand for something like *more than 10, more ... than the number of girls*, etc. These are *DET*s by our criterion, but not the single *more*.

This is an objection to (Q) that must be taken seriously. It involves (i) finding a semantic distinction between the quantifiers **more** and, say, **most**, which explains why one but not the other is a *DET* denotation; (ii) the analysis of 'there are'-sentences; (iii) the semantics of words like *more*. These matters will be taken up in Section 2.2.

2.1.2 Determiner interpretations as generalised quantifiers

Following Montague, we interpreted *DET*s as *functions* from subsets of the universe M to sets of such subsets. From now on, however, we return to the generalised quantifier framework of Section 1, where quantifiers on M are *relations* between subsets of M. Thus, to each n-place function $\mathbf{D}$ from $(P(M))^n$ to $P(P(M))^n$ we associate the following $(n+1)$-ary quantifier on M:

$$\mathbf{Q}_M A_1 \dots A_n B \Leftrightarrow B \in \mathbf{D}(A_1, \dots, A_n).$$

In what follows, *DET* interpretations will be such monadic quantifiers on the universe.

The functional interpretation of *DET*s emphasises similarity of structure between syntax and semantics, which is one of the characteristics of Montague Grammar. From the present semantic perspective, however, relations turn out to be easier to work with. But keep in mind that the relational approach increases the number of arguments by one: n-place *DET*s will denote $(n+1)$-ary quantifiers (so far we have only seen 1-place *DET*s, but cf. 2.2). It should also be noted that for *some* semantic issues, the functional framework seems more natural; cf. [Keenan and Moss, 1985].

Terminological Remark: The use of words 'determiner' and 'quantifier' is rather shifting in the literature. Here, the idea is to use 'determiner' and '*DET*' *only* for syntactic objects, and 'quantifier' *only* for semantic objects. The extension of 'quantifier' was given in Section 1.4, and a criterion for *DET*-hood at the beginning of 2.1.

2.1.3 Determiners as constants

In a Montague-style model $\mathbf{M} = \langle M, \| \ \| \rangle$, *DET*s are on a par with expressions of other categories. Nothing in principle prevents, for example, that a determiner D is interpreted as **every** in one model and as **most** in another. But there is usually no point in allowing this generality. Moreover, there is a clear intuition, I think, that determiners are *constants*. We therefore lay down the following *methodological postulate*:

(*MP*) *Simplex DETs are constants: each one denotes a fixed quantifier* (modulo, of course, lexical ambiguity, vagueness, etc.; cf. 2.4).

(*MP*) allows us to dispense with the interpretation function for (simples) *DET*s and to resume the notation from 1.4, using boldface letters for quantifiers: Q denotes **Q**, *most* denotes **most**, *some* denotes **some**, etc.

What about complex *DET*s? In case such a *DET* contains a non-constant expression, there seems to be a choice. We can either persist in treating them as constants, or let their interpretation depend on the interpretation of the non-constant expressions occurring in them. To take a simple example, consider *some red*. This expression *can* be construed as giving an *NP* when applied to an N, thus *can* be classified as a *DET* by our criterion. As a constant, it would denote the quantifier defined by

$$\textbf{some red}_M AB \Leftrightarrow A \cap B \cap \{a \in M : a \text{ is (in fact) red}\} \neq \varnothing,$$

for each universe M. As an expression consisting of a constant and a non-constant symbol, i.e. of the form *some* P, it is interpreted in a model **M** as

$$\|\textit{some } P\|AB \Leftrightarrow A \cap B \cap \|P\| \neq \varnothing.$$

Given **M**, this is a quantifier on M, but the expression does not denote a fixed quantifier on each universe.

No doubt many readers will find the latter option more natural, but we need not take a stand on this methodological issue here. Our model-theoretic machinery provides adequate semantic objects for both cases, quantifiers, and quantifiers *on* universes, respectively.

Note, however, that our decision to treat simplex *DET*s as constants does not necessarily imply that they are *logical* constants. It can be argued that logicality requires a lot more; this theme will be resumed in 3.4 and 4.4 (cf. also [Westerståhl, 1985a]). For example, the quantifier **some red** defined above is not logical, the reason being that it violates the condition *ISOM* from 1.4.

In Appendix B we will indicate what happens if the postulate (*MP*) is dropped.

2.1.4 Global vs. local perspective

To study quantifiers from a *global* perspective means to concentrate on properties which are *uniform* over universes. A typical example is first-order definability: **Q** is first-order definable if there is some first-order sentence which defines it on *every* universe. Sometimes, however, it is natural to take a *local* viewpoint: fix a universe M and restrict attention to quantifiers on M. Then other definability notions become interesting as well, involving parameters from M in an essential way.

Our perspective here will be predominantly global. The main reason for this is that global definitions and results are more general: they usually have an immediate 'local version'. The converse, however, does not hold. Quantifiers from a local perspective are studied extensively in [Keenan and Stavi, 1986]. Some of their results will be reviewed in Section 4.6.

2.2 The Interpretation of Determiners

The basic quantifier postulate (Q) from 2.1.1 can be split into a syntactic and a semantic part as follows:

(Q_{syn}) *Quantifier expressions are DETs.*

(Q_{sem}) *DETs denote* $(n+1)$*-ary quantifiers,* $n \geqq 1$.

In contrast with standard predicate logic, there are no unary quantifiers on this approach. And although some binary *DET* denotations (e.g. Montague's **every, a, the**) are definable in standard predicate logic, others are not: we saw in 1.7 that **most** is an example. Consequently, *EL* is inadequate for formalising even the pure quantificational part of natural languages.

However, (Q) is not yet quite satisfactory. In particular, we need to account for the apparent counter-examples mentioned in 2.1.1, III. Nothing so far precludes **more** from being a *DET* denotation.

The starting-point of a systematic study of natural language quantification was the isolation, in [Barwise and Cooper, 1981], and independently in [Keenan and Stavi, 1986] (although the latter paper was published much later, they were written at about the same time), or a purely model-theoretic property characteristic of those quantifiers that are *DET* denotations. This is the property of *conservativity*, defined below (Barwise and Cooper used a different terminology, in terms of an NP denotation *living on* a given set). Actually, the property (and the term) first appeared in [Keenen, 1981], but in the two first-mentioned papers it was proposed as a significant semantic universal for determiners (although with rather different motivations; cf. below).

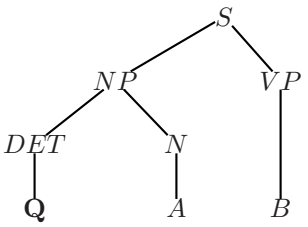
2.2.1 Conservativity

A binary quantifier **Q** is called *conservative* if the following holds:

(*CONSERV*) *for all* M *and all* $A, B \subseteq M$, $\mathbf{Q}_M AB \Leftrightarrow \mathbf{Q}_M A\ A \cap B$.

It is easily checked that **most** is conservative, but **more** is not. As we will see in 2.4, practically all English *DET*s denote conservative quantifiers (a few possibly doubtful cases will be noted).

CONSERV gives the *first* argument of **Q** a privileged role: only the part of B which is common to A matters for whether $\mathbf{Q}_M AB$ holds or not. This semantic difference between the arguments A and B matches the syntactic difference between the corresponding expressions:



Conservativity is a very fruitful postulate, as well be seen in Sections 3 and 4. Still, one may ask what, if any, is the idea or intuition behind it. As for Barwise and Cooper, they seem to regard it mainly as a successful empirical generalisation. Keenan and Stavi, on the other hand, give an interesting theoretical motivation: they prove that, on a given (finite) universe M, the conservative quantifiers on M are precisely those which can be generated from certain initial quantifiers by means of a few natural closure operations; an exact statement (and proof) will be given in Section 4.6. Yet another motivation, discussed in [Westerståhl, 1985a], is that *CONSERV* is related to the notion of *restricted domains of quantification*: an *NP* 'restricts' the universe to the denotation of the N; this will be formulated in Section 3.2.

CONSERV resolves the first doubt concerning (Q) expressed in 2.1.1, III. We still have to deal with 'there are'-sentences and with the semantics of *more*.

2.2.2 'there are'-sentences

Consider sentences such as

(1) There are no flowers,

(2) There are many patients waiting outside,

(3) There are some philosophers who like logic,

(4) There are a few errors in the text.

Without commitment to their syntactic form, let us write such sentences

(5) *There are* $\mathbf{Q}_M A$,

where **Q** is the quantifier denoted by the *DET* and A is the set contributed, in a model **M**, by the expression following the *DET*.[16] There are in fact two questions here. The first is to interpret quantified sentences of the form (5) in a way consonant with the basic postulate (Q). The second concerns the fact that certain *DET*s do not fit in (5): *all, most, not all*, for example. Is there a semantic explanation for this phenomenon?

[16]The 'hybrid' form (5) is used in order to avoid discussion of the syntactic structure of "there are"-sentences. This structure is quite varied, as already (1)–(4) indicate, and there may be divergent opinions about it, but it still seems that (5) is *semantically* adequate in a large number of cases.

We shall review the answers proposed by Barwise and Cooper to both of these questions, first, because they show a way to handle 'there are'-sentences, and second, because this case can serve as a model of the kind of linguistic explanation one may expect from the present theory of quantifiers.

The first proposal is simple: interpret (5) as

(6) $\mathbf{Q}_M AM$.

This interpretation *works* in the sense that it gives (1)–(3) the right truth conditions. Moreover, one can argue that it accounts for the idea that the phrase 'there are' serves to ascribe *existence*, i.e. the property that everything in the universe has, to the rest of the sentence.

But why are some choices of Q apparently forbidden in (5)? First, a definition. Call a *DET strong*, if its denotation, as a binary relation, is either reflexive or irreflexive; otherwise the *DET* is *weak*. Now observe that the *DET*s that fit in (5) are weak, whereas the exceptions are strong.[17] This is still no explanation, but it is a *fact* which may point to one. The next move is theoretical: we *prove* in our theory that (6) is equivalent to

(7) $\mathbf{Q}_M AA$;

this is actually an immediate consequence of CONSERV. It follows that

If Q is strong, (5) is either trivially true or trivially false.

Thus, the connection between the strong/weak distinction and our problem has not merely been *described*; it has been *explained*, given the plausible assumption that it is in general 'strange' to utter trivial truths or falsities.

This simple but instructive model of explanation shows the typical interplay between linguistic facts, theoretical concepts, and results in the theory. Here the results used were quite trivial,but this may not always be the case.

Let me hasten to add that the above by no means exhausts the many interesting problems connected with 'there are'-sentences. Moreover, Keenan and Stavi [1986] argue against the explanation in terms of the strong/weak distinction; the propose another semantic characterisation of the relevant class of determiners (a detailed discussion of these matters can be found in [Keenan, 1989]). But it is the *type* of explanation that I have tried to illustrate here.

2.2.3 $(n+1)$-ary conservative quantifiers

Now, what about *more*? We noted that

(8) There are more P than Q

(9) *more* $x(Px, Qx)$.

[17] Actually, *most*, as we have interpreted it, is not reflexive, since $\mathbf{most}_m AA$ is false when $A = \varnothing$. One remedy is to redefine it for this argument.

Observe further that *more P than Q* is very naturally considered as *NP*, obtained by applying the 2-place *DET more ... than* to two Ns, and typically occurring in sentences such as

(10) More men than women voted for Smith.

(10) means that the number of men who voted for Smith is greater than the number of women who voted for Smith. So *more ... than* denotes a *ternary* quantifier:

$$\textbf{more ... than}_M A_1 A_2 B \Leftrightarrow |A_1 \cap B| > |A_2 \cap B|.$$

Other examples of such ternary quantifiers are

$$\textbf{fewer ... than}_M A_1 A_2 B \leftrightarrow |A_1 \cap B| < |A_2 \cap B|,$$
$$\textbf{as many ... as}_M A_1 A_2 B \Leftrightarrow |A_1 \cap B| = |A_2 \cap B|.$$

We now see that (8) can be written as a generalisation of (5) to ternary quantifiers:

(11) *There are* $\mathbf{Q}_M A_1 A_2$.

Furthermore, (11) can be interpreted on exactly the same principle as (5), namely, as

(12) $\mathbf{Q}_m A_1 A_2 M$.

For example, if P denotes A and Q denotes B, the interpretation of (8) is

$$\textbf{more ... than}_M ABM,$$

which is equivalent to

$$|A| > |B|$$

i.e. to

$$\textbf{more}_M AB,$$

as predicted. So the previous analysis of 'there are'-sentences with binary quantifiers extends naturally to ternary (in fact, $(n+1)$-ary) quantifiers. (The reader might wish to ponder whether the characterisation in terms of the strong/weak distinction also generalises; cf. [Keenan, 1989]).

Finally, the notion of conservativity also extends to $(n+1)$-ary quantifiers: the set to which the VP denotation can be restricted is then the *union* of the n denotations. We get the following general version of *CONSERV* for $(n+1)$-ary quantifiers:

(*CONSERV*) *For all M and all $A_1, \ldots, A_n, B \subseteq M$,*
$\mathbf{Q}_M A_1 \ldots A_n B \leftrightarrow \mathbf{Q}_m A_1 \ldots A_n (A_1 \cup \ldots \cup \_n) \cap B$.

It is easily verified that *more ... than, fewer ... than, as many ... as* are all conservative, in contrast with the binary operator **more**.

In conclusion, our findings about the use of *more* do not contradict the basic idea (Q), on the contrary, they support it. A final formulation of this idea goes as follows (cf. the beginning of 2.2):

(Q_{syn}) *Quantifier expressions are DETs.*

(Q_{sem}) *n-place DETs denote $(n+1)$-ary conservative quantifiers, $n \geqq 1$.*

We should perhaps note that there are other uses of *more* in determiners, for example, *more than ten* or *six or more*. These are ordinary (complex) 1-place *DET*s, and denote binary conservative quantifiers, just as (Q_{sem}) predicts (cf. also 2.4.7).

2.3 Subject-Predicate Logic

As in Montague Grammar, Barwise and Cooper use an intermediate logical language, called $L(GQ)$, into which a fragment of English is translated. $L(GQ)$ has two unusual features;

(i) Quantified sentences have $NP - VP$ form (subject-predicate form).

(ii) Quantifier symbols are not used as variable-binding operators.

The well-formed expressions in $L(GQ)$ are of two kinds: *formulas* and *set terms*. A set term is either a unary predicate symbol or an expression of the form

$$\hat{x}[\psi],$$

where x is a variable and ψ a formula; in models, set terms denote subsets of the universe. Variable-binding is done with the *abstraction operator* $\hat{}$. Quantifier symbols are (certain) 1-place *DET*s and quantified formulas are of the form

(*) $D(\eta)(\delta)$,

where D is a *DET* and η, δ are set terms. There are the usual atomic formulas, plus formulas of the form $\eta(t)$, where η is a set term and t an individual term, and the formulas are closed under sentential connectives. *DET*s are interpreted as binary conservative quantifiers; the truth condition for (*) in a model is then obvious.

The result is that logical form in $L(GQ)$ corresponds more closely to syntactic form in the fragment than usual. (*) can be said to have $NP - vP$ form with $D(\eta)$ as the NP and δ as the VP (the formation rules actually give (*) this structure). Another pleasant feature is that some unnecessary uses of bound variables are avoided. For example,

(1) Some boys run

is translated

(1′) *some*(*boy*)(*run*)

instead of the usual

(1″) $\exists x(boy(x) \wedge run(x))$.

the example also shows that certain unnecessary sentential connectives in the standard formalisation are avoided. In more complex cases, e.g. with transitive verbs or relative clauses, $L(GQ)$ must introduce variables and connectives (though English often can avoid them): consider

(2) Most women who love Harry have a cat,

(2′) *most* ($\hat{x}$[*woman*(x) $\wedge$ *love*(x, *Harry*)])($\hat{x}$[*a*(*cat*)($\hat{y}$[*have*(x, y)])]),

(2″) *most* x(*woman*(x) $\wedge$ *love*(x,*Harry*),$\exists$*y*(*cat*(y) $\wedge$ *have*(x, y))).

These examples should make it plausible that there is no deep difference between $L(GQ)$ and the standard language for generalised quantifiers as in 1.4. In fact, they are even syntactically intertranslatable in a rather obvious way. Still, quantified formulas in $L(GQ)$ have subject-predicate form. It is hard to avoid the conclusion that the importance of the issue of whether subject-predicate form occurs in logic has been greatly over-estimated, from Russell and onwards.

2.4 Some Natural Language Quantifiers

A quantifier **Q** will be called a (*simple*) *natural language quantifier*, if it is denoted by some (simplex) natural language *DET*.

This notion is somewhat loose,but its serves our purposes. A more exact specification would presuppose, among other things, (i) that the class of *DET*s has been more precisely delimited; (ii) that it has been decided how to treat complex non-logical *DET*s (2.1.3); (iii) that a global or a local perspective has been chosen (2.1.4). We may think of the notion of a natural language quantifier as having various *parameters*, which can be set at different values. It turns out that, for many of the things we shall have to say about natural language quantifiers, the value of these parameters is immaterial. This is why the above 'loose' notion is useful. And in other cases, we will indicate how a particular observation on natural language would depend on different parameter settings.

To take a first and crude example, consider the assertion that *not all binary quantifiers are natural language quantifiers*. From a global perspective, or from a local perspective with a given infinite universe M, this is true for cardinality reasons: there are uncountably many binary quantifiers (on M), but at most countably many natural language quantifiers. But, even from a finite local perspective, the assertion is true for another reason, namely, the conservativity universal (e.g. **more** or **more**$_M$ is not a natural language quantifier). The other parameter settings are

clearly irrelevant her, so the assertion is true however the parameters are set. An example of an assertion whose truth does depend on the parameters is this: *All natural language quantifiers satisfy ISOM*. We will see in section 3.3 that this is in fact a candidate for a quantifier universal, but only under a certain delimitation of the class of *DET*s.

In the remainder of this section, I will present a list of examples of natural language quantifiers. Some of them will be used later on, but the list is also intended to give the reader a feeling for the perhaps surprising richness of the class of natural language quantifiers.

The method is simply to list the various English *DET*s, together with their semantic interpretations (when these are not obvious). The *DET*s are selected by using the criterion for *DET*-hood in Section 2.1 as liberally as possible, but with some 'common sense' (standard co-occurrence criteria for constituenthood, etc.). Thus I will be listing *possible DET*s — there may be syntactic, semantic, or methodological reasons for discarding several of them from a more definitive list. In fact, some such reasons will be discussed in what follows.

The main sources for the list that follows are [Keenan and Stavi, 1986] and [Keenan and Moss, 1985]. The reader is referred to these works for further examples, and for detailed arguments that most of the expressions listed really belong to the category *DET*.

2.4.1 Some simplex DETs

(1) *all, every, each, some, a, no, zero, most*

(2) *both, neither*

(3) *one, two, three, . . .*

(4) *many, few, several, a few*

(5) *the*

(6) *this, that, these, those*

(7) *more . . . than, fewer . . . than, as many . . . as*

Here are some interpretations, a few of which have already been given

$$\begin{aligned}
&\mathbf{all}_M AB \Leftrightarrow \mathbf{every}_M AB \Leftrightarrow \mathbf{each}_M AB \Leftrightarrow A \subseteq B,\\
&\mathbf{some}_M AB \Leftrightarrow \mathbf{a}_M AB \Leftrightarrow A \cap B \neq \varnothing,\\
&\mathbf{no}_M AB \Leftrightarrow \mathbf{zero}_M AB \Leftrightarrow A \cap B = \varnothing,\\
&\mathbf{most}_M AB \Leftrightarrow |A \cap B| > |A - B|,\\
&\mathbf{both}_M AB \Leftrightarrow \mathbf{all}_M AB \& |A| = 2,\\
&\mathbf{neither}_M AB \Leftrightarrow \mathbf{no}_M A \& |A| = 2,\\
&\mathbf{one} = \mathbf{some},\\
&\mathbf{two}_M AB \Leftrightarrow |A \cap B| \geqq 2,\\
&\mathbf{three}_M AB \Leftrightarrow |A \cap B| \geqq 3, \ldots
\end{aligned}$$

So **n** is interpreted as **at last n** here, although it can be argued that it sometimes means **exactly n**. As for (4)–(6), cf. 2.4.2–6 below. The denotation of the 2-place *DET*s in (7) were given in 2.2.3.

2.4.2 Vague DETs

Vagueness in the sense of the occurrence of *borderline cases* (in some suitable sense) pertains to *DET*s as well as to other expressions. We do not incorporate a theory of vagueness here, but choose idealised precise versions instead.

Two examples of vague *DET*s are *several* and *a few*. Here one may, following Keenan and Stavi, stipulate that

several = three,
a few = some.

2.4.3 Context-dependent DETs

The *DET*s *many* and *few* are not only vague but also context-dependent in the sense that the 'standard of comparison' may vary with the context. For example, in

(8) Many boys in the class are right-handed,

(9) Lisa is dating many boys in the class,

some 'normal' standard for the least number considered to be many is used,but probably different standards in the two cases. Even within one sentence different standards may occur, as in the following example (due to Barbara Partee):

(10) Many boys date many girls.

Other, complex, *DET*s with a similar behaviour are, for example,

a large number of, unexpectedly few, unusually many.

Westerståhl [1985a] discusses various interpretations of *many*. Basically, there are two possible strategies. Either one excludes this type of *DET*s from extensional treatments such as the present one (this is what Keenan and Stavi do), or one tries to capture what *many* might mean in a *fixed* context (this is the approach of Barwise and Cooper). Here are some suggestions for the second strategy:

$$\begin{aligned}
&\mathbf{many}^1_M AB \Leftrightarrow |A \cap B| \geqq k|M| && (0 < k < 1),\\
&\mathbf{many}^2_M AB \Leftrightarrow |A \cap B| \geqq k|A| && (0 < k < 1),\\
&\mathbf{many}^3_M AB \Leftrightarrow |A \cap B| \geqq (|B|/|M|)|A|.
\end{aligned}$$

$\mathbf{many}^1$ relates the standard to the size of the universe: in a universe of 10, 5 may be many, but not in a universe of 1000. $\mathbf{many}^2$ is a *frequency* interpretation: the number of As that are B, compared to the total number of As, is at least great as

a 'normal' frequency of Bs, given by k. In both cases, k has to be supplied by the context. But in $\mathbf{many}^3$, the 'normal' frequency of Bs is just the actual frequency of Bs in the universe.

Notice that $\mathbf{many}^1$ and $\mathbf{many}^3$ make essential reference to the *universe* of the model. As we shall see, this is in contrast with most other natural language quantifiers. Also notice that $\mathbf{many}^3$ is *not conservative*. Since the conservativity universal is so central, this observation gives a (methodological) argument for discarding $\mathbf{many}^3$ as an interpretation of *many*.

As for *few*, we may simply interpret it as *not many*.

2.4.4 Ambiguous DETs

Ambiguity in the sense of a small number of clearly distinguishable meanings of a *DET* is another phenomenon than context-dependence. We have already noted that the *DET*s *one, two, three, ...* may be ambiguous with respect to the quantifiers **at least n** and **exactly n**. Another possibly ambiguous *DET* is *most*: it can be argued that aside from the interpretation we have given, *most* can also mean something like *almost all*; cf [Westerståhl, 1985a].

The fact that certain *DET*s may be ambiguous is not a problem in the present context, as long as we make sure to include each of their interpretations among the natural language quantifiers.

2.4.5 Pronominal DETs

Most 1-place *DET*s can occur without their N arguments, as was noted in 2.1.1. Such *DET*s may be called *pronominal*. The natural analysis of sentences with pronominally occurring *DET*s is that the argument (or the set it denotes) is given by the context. So

All cheered

is interpreted as

$$\mathbf{all}_M X \| \mathit{cheered} \|,$$

where the set X is provided by the context. The use of such *context sets* is studied further in [Westerståhl, 1985b].

The only non-pronominal 1-place *DET*s encountered so far are, as the reader can check,

a, every, no, the.

Moreover, *DET*s taking two or more arguments are *never* pronominal, it seems.

Note that the pronominal *all* and the non-pronominal *every* denote the same quantifier. So pronominality is not a semantic property of *DET*s in the present framework.

2.4.6 Definites

By the *simple definites* we shall understand here

(i) the definite article *the*,

(ii) the *simple possessives*, like *John's, Susan's, my, his, their*,

(iii) the demonstratives: *this, that, these, those*.

We have already given an interpretation for *the*:

$$\mathbf{the}_M AB \Leftrightarrow \mathbf{all}_M AB\&|A| = 1.$$

This is the *singular the*, as in

(11) The boy is running.

For a sentence like

(12) The boys are running

we must use instead

$$\mathbf{the}^{\mathrm{pl}}_M AB \Leftrightarrow \mathbf{all}_M AB\&|A| > 1.$$

Thus *the* is ambiguous on this analysis. Demonstratives can be interpreted similarly; there we have singular and plural forms and thus no ambiguity. but the simple possessives exhibit the same ambiguity as *the*:

(13) John's car is clean,

(14) John's cars are clean

can be interpreted, respectively, with the quantifiers

$$\mathbf{John's}_M AB \leftrightarrow \mathbf{all}_M P_{\mathrm{John}} \cap AB\&|P_{\mathrm{John}} \cap A| = 1,$$
$$\mathbf{John's}_M AB \leftrightarrow \mathbf{all}_M P_{\mathrm{John}} \cap AB\&|P_{\mathrm{John}} \cap A| > 1,$$

where P_{John} is the st of things possessed by John; a possession relation is then supposed to be given in the model. there are also *relational* uses of possessives, where the relation is given explicitly, as in

(15) John's friends are nice.

Here it is doubtful whether *John's* applies to an N, and thus whether it is a *DET* in our sense. (In any case, the truth condition for sentences like (15) can be given by

$$\mathbf{John's}^{\mathrm{pl}}_M RB \Leftrightarrow \mathbf{all}_M R_{\mathrm{John}} B\&|R_{\mathrm{John}}| > 1,$$

where R is a binary relation on M and $R_a = \{b \in M : Rab\}$ — here we have a generalised quantifier of type $\langle 2, 1\rangle$.)

We see that the definites come with a *number condition*, concerning the number of elements in a certain set. It is also possible to let sentences with definites *presuppose* that the number condition is satisfied, instead of making them false when it isn't, as we did above. This could be effected by extending the model-theoretic framework to allow *partial* quantifiers **the**, **the**$^{\text{pl}}$, **John's**, **John's**$^{\text{pl}}$ would then be *undefined* when the number condition is not met. We return to this in 3.7.

2.4.7 Complex DET*s with definites*

There are several ways to construct complex *DET*s with definites in English, in particular with partitive constructions. I will present a rather uniform way of interpreting such *DET*s. The starting-point is the observation that one function of the simple definites is to indicate the occurrence of *context sets* (cf. 2.4.5). For simple possessives, this is usually the set of things possessed by the individual (it may also be a subset of this set). But also *the* and the demonstratives need context sets to make the interpretation come out right. For example, in (11) or (12) we are usually *not* talking about the set of all boys in the universe M (as the interpretations given in 2.4.6 would have us believe), but a context-given subset of it (in the singular case, this set has one element).

Consider sentences (with *DET*s as indicated) like

(16) *Some of the seven* men survived,

(17) *Most of John's few* books were stolen.

We interpret these on the following scheme:

(18) $(\mathbf{Q}_1$ **of Def** $\mathbf{Q}_2)\ BC \Leftrightarrow \mathbf{Q}_1 X \cap BC \& \mathbf{Q}_2 X \cap BM$,

where $\mathbf{Q}_1, \mathbf{Q}_2$ are quantifiers and *Def* is a simple definite with X as associated context set (the subscript 'M' is omitted for readability). note that the second conjunct in (18) can be written, as in 2.2.2,

There are $\mathbf{Q}_2 X \cap B$,

expressing the condition that, in (17), John's books were few, and, in (16) that the set of men under consideration has (exactly?) seven elements.

Some other constructions with definites can be obtained as special cases of (18). We define

(19) $\mathbf{Def}\ \mathbf{Q}_2)BC \Leftrightarrow$ (**all of Def** $\mathbf{Q}_2)BC$,

(20) $(\mathbf{Q}_1$ **of Def**$)BC \Leftrightarrow (\mathbf{Q}_1$ **of Def all**$)BC$
$\Leftrightarrow \mathbf{Q}_1 X \cap BC$ (by (19) with $\mathbf{Q}_2 = \mathbf{all}$),

(21) $\mathbf{Def}BC \Leftrightarrow$ (**all of Def**$)BC$
$\Leftrightarrow \mathbf{all} X \cap BC$ (by (20) with $\mathbf{Q}_1 = \mathbf{all}$).

(19) takes care of complex *DET*s such as

the five, these few, John's several, etc.

(20) deals with partitives such as

some of Susan's, many of these, at least five of the, etc.

And (21) returns to the simple definites: the truth conditions are essentially the same as in 2.4.6, except that context sets are mentioned.

(18)–(21) can be seen to give the right truth conditions for sentences of these forms, *except* that we have, for readability, omitted the number conditions belonging to these interpretations: in (18) and (20) a *plural condition*, i.e. that $|X \cap B| > 1$, should be added, and in (19) and (21) the cases with singular and plural conditions should be distinguished (syntactically they are distinguished by the singular or plural form of the N denoting B).

More complicated *DET*s with definites can be treated along similar lines. For example, there are *DET*s which quantify over the *possessor* a in a simple possessive

$$a\text{'}\mathbf{s}BC \Leftrightarrow \mathbf{all}P_a \cap BC$$

(we continue to leave out the number condition, and assume for simplicity, in the rest of this subsection, that everything is in the plural). One example is with *DET*s like

some students', most boys', several girls', etc.,

as in

(22) Some students' books were stolen.

The interpretation of these *DET*s is given by

(23) $(\mathbf{Q}_1 A\text{'}\mathbf{s})\mathbf{C} \Leftrightarrow \mathbf{Q}_1 A\{a \in M : a\text{'}\mathbf{s}BC\}$.

Another example is with *iterated* definites. Here is one scheme, which generalises (20):

(24) $(\mathbf{Q}_1$ **of Def** $A\text{'}\mathbf{s})BC \Leftrightarrow \mathbf{Q} - 1X \cap A\{a \in M : a\text{'}\mathbf{s}BC\}$

(we could have generalised (18) similarly, but examples of this form seem rare). This covers *DET*s like

most of the students', some of these boys', three of John's cars', etc.

It could be argued that a sentence like

(25) Most of the students' books were stolen

is ambiguous; then (24) gives the sense where *most* takes *students* as argument, whereas the sense where it takes *books* as arguments is given by

(24) $(\mathbf{Q}_1$ **of DEF** A'**s**$)BC \Leftrightarrow$ **all**$X \cap A\{a \in M : (\mathbf{Q}_1$ **of** a'**s**$)BC\}$.

As before, if *the* is replaced by *John's* in (25), $X = P_{\mathrm{John}}$ (or a subset of it) n (24) and (26). Also as before, we get *DET*s like

the students' those boys', Susan's cars', etc.

as a special case of (24):

(27) $($**Def** A'**s**$)BC \Leftrightarrow ($**all of Def**A**'s**$)BC$,

and similarly for *DET*s like

the five students', those few boys', Susan's two cars', etc.

We have given uniform truth conditions for a number of sentences with complex *DET*s by proposing a semantics for the *DET constructions* involved there. This is one task of a theory of natural language quantification. Another is to describe and if possible explain the *restrictions* that often belong to such constructions (cf. 2.2.2).

Consider, for example, the construction in (18). One can see that only *pronominal DET*s can be in the $\mathbf{Q}_1$ position here. As for the **Def** position, the definites, and no others, will work. And there are restrictions on $\mathbf{Q}_2$ too: e.g. *most, all, every, no, some* sound strange here. This last restriction can actually be explained by combining the Barwise and Cooper explanation of the restrictions on 'there are'-sentences (2.2.2) with the *plural condition* holding for (18): the exceptions will then once more be those quantifiers making the truth condition trivial. This and other restrictions at work here are discussed further in [Westerståhl, 1985b].

There is one notable feature of the constructions with definites given here: although the analysis is compositional, it does not use the quantifiers taken to interpret the simple definites in 2.4.6. The function of simple definites was merely to provide context sets. If our analysis is viable, it opens the possibility to leave out the definites from the class of *DET*s, i.e. to treat them as not denoting quantifiers. This move has in fact been viewed desirable for independent reasons which I will not discuss here. My point is merely that such a move can be accommodated in the present quantifier framework.

Likewise, it is not strictly necessary to regard the constructions in this subsection as giving new *DET*s and thereby new natural language quantifiers. Instead, the definitions (18)–(21), (23)–25), (27) *could* be seen as uniform truth conditions for *sentences* involving (among other things) quantifiers $\mathbf{Q}_1$ and $\mathbf{Q}_2$, but not as defining new quantifiers. the class of natural language quantifiers will then become correspondingly smaller.

If, on the other hand, these constructions are regarded as quantifier definitions, it should be noted that they always yield *conservative* quantifiers, provided $\mathbf{Q}_1$ and $\mathbf{Q}_2$ are conservative.

Clearly we have merely scratched the surface of the many problems pertaining to the analysis of definites, possessives, partitives, etc. It seems, however, that the present quantifier framework can be applied quite fruitfully to these well known linguistic questions; cf. for example [Keenan and Stavi, 1986; Partee, 1984a; Partee, 1984b; Thijsse, 1983].

2.4.8 Numerical DETs

There are many variations of the simplex numerical *DET*s *one, tow, three*, ..., e.g.

> *at least five, at most five, exactly five, five or more, between five and ten, more than five, fewer than five, infinitely many, at most finitely many, an even number of, an infinite number of, every other, every third, around ten, almost ten, nearly ten, approximately ten, ...;*

the interpretations are more or less obvious. A particular group of numerical expressions is

> *half, more than half, less than half, at least half, not more than half, two thirds, at least two thirds, ...*

These are not really *DET*s by our criterion (they don't apply to Ns), but if a phrase of the form *of Def* is appended to hem (after *half*, the *of* is optional), the resulting expressions are quite similar to those in (20): *more than half of the, two thirds of John's, not more than half of these*, ... the interpretation give in (20) fits well here,but to use it we must have suitable quantifiers $\mathbf{Q}_1$ available. Thus, it seems reasonable, even if the above expressions are not *DET*s, to include the quantifiers

> **at least m/n** $AB \Leftrightarrow |A \cap B| \geqq m/n|A|$

$(n > m > 0)$ among the natural language quantifiers (Boolean combinations of these will then give the other quantifiers needed here).

2.4.9 Comparative DETs

The words *more, fewer, less*, ... can be used in *DET*s for comparison with a fixed number or proportion, as in 2.4.8. We also have the 2-place simplex *DET*s *more ... than, fewer ... than,* etc. Some complex variants of these are

> *more than twice as many ... as, less than half as many ... as, etc.*

Keenan and Stavi discuss other comparative *DET*s, e.g. those in

(28) *More male than female* students stayed home,

(29) *More* students attended *than stayed home*,

(30) *More* students attended *than teachers who stayed home*;

the respective 1-place *DET*s are italicised. That they are putative *DET*s follows by our criterion (nothing prevents a 1-place *DET* from being syntactically discontinuous!). However, it is also possible to analyse (28)–(30) with the 2-place *more . . . than*: rewrite them as

(28′) *More* male students *than* female students stayed home,

(29′) there are *more* students who attended *than* students who stayed home,

(30′) there are *more* students who attended *than* teachers who stayed home.

The last two 'there are'-sentences are then treated as in 2.2.3.

These examples illustrate nicely that more than one structural analysis of an NP is often possible. Since no semantic ambiguity is involved here, one would like to make a choice. For a further illustration, consider

(31) *More* men *than* women voted for Smith,

(31′) *More* men *than women* voted for Smith.

(31) uses *more . . . than*, whereas (31′) uses the 1-place *more than women*. but this latter *DET* is not conservative, as one easily sees, so we have a good reason to prefer (31). The *DET*s in (18)–(30), on the other hand, are all conservative. For example,

$$\begin{aligned}&\textbf{more than stayed home}_M AB \Leftrightarrow\\ &\quad\Leftrightarrow |A \cap B| > |A \cap \|\textit{stayed home}\|\|\\ &\quad\Leftrightarrow |A \cap (A \cap B)| > |A \cap \|\textit{stayed home}\|\|\\ &\quad\Leftrightarrow \textbf{ more than stayed home}_M AA \cap B.\end{aligned}$$

Still, there are reasons to prefer (28′)–(30′). One is that they are simpler and more uniform. Another will be given in Section 3.3.

Keenan and Stavi also consider comparatives with definites, such as

more of John's than of Susan's, fewer of the male than of the female, etc.

These can be dealt with, if one wishes, by combining the simplex 2-place comparatives with the treatment of definites in 2.4.6 an d2.4.7; we omit details.

2.4.10 "Only"

Consider the sentence

(32) Only women voted for Smith.

If *only* is a *DET* here, its interpretation is

$$\textbf{only}_M AB \leftrightarrow B \subseteq A.$$[18]

This is *not* a conservative quantifier (indeed, $\textbf{only}_M AA \cap B$ is trivially true for all A, B). So let us look for alternatives. Now, *only* can modify many other things besides Ns, e.g. *NP*s:

(33) Only Susan voted for Smith.

An alternative analysis is then to treat *women* in (32) as a full em NP (a 'bare plural'); then *only* is not a *DET* at all.

there are also complex *DET*s with *only*. Consider the following example (essentially from Keenan and Stavi):

(34) Only liberal students voted for Smith.

This sentence is three ways ambiguous: (i) as an answer to 'Which students voted for Smith?'; (ii) as an answer to 'Which liberals voted for Smith?'; and (iii) as an answer to 'Who voted for Smith?'. Writing (34) in the form *only* ABC, we can represent its three meanings as

(i) *only* $ABC \Leftrightarrow B \cap C \subseteq A$,

(ii) *only* $ABC \Leftrightarrow A \cap C \subseteq B$,

(iii) *only* $ABC \Leftrightarrow C \subseteq A \cap B$.

There are various possibilities here. One is to treat *only* as a 2-place *DET* with three possible interpretations, as in (i)–(iii). One readily verifies that (i) and (ii), but not (iii), are conservative. Or, if one wants to analyse (34) with a 1-place *DET*, we have, in case (i),

$$\textbf{only liberal}_M AB \Leftrightarrow A \cap B \subseteq \|\textit{liberal}\|;$$

in case (ii),

$$\textbf{only . . . students}_M AB \Leftrightarrow A \cap B \subseteq \|\textit{student}\|$$

(but *only . . . students* isn't really a *DET* since it applies to an adjective); and in case (iii) the ordinary *only*, as in (32). Again, the first two are conservative, but not the third.

Only can also combine with numerical expressions, as in

(35) Only five students voted for Smith.

[18] One may argue that (32) also says that *some* women voted for Smith. We ignore the possible existence implications of *only* here, but they could easily be added without affecting the discussion.

This time, there is no analysis with a 2-place *DET*, and there are just two possible meanings: (i) as an answer to 'How many students voted?'; and (ii) as an answer to 'How many voted?'. So, writing (35) as *only five* AB, we get

(i) *only five* $AB \Leftrightarrow$ **exactly five**$_M AB$,[19]

(ii) *only five* $AB \Leftrightarrow$ **exactly five**$_M AB \& B \subseteq A$.

In case (ii), *only five* would be a non-conservative *DET*, but it is more natural to treat *only* as an NP-modifier here. In case (i), on the other hand, *only five* works fine as a conservative *DET*. Here one would like to see a uniform treatment of *DET*s of the form

(36) *only* Q;

we have already seen that *only* 'transforms' n into *exactly* n, but when Q is a definite, things get more complicated, as the reader can check by considering the example

(36) Only John's students voted for Smith

(three possible readings). Also, one would like to explain the restrictions on Q in (36). For example, *a few, between five and ten, around ten* are fine, but not *several, all, most*.

These are just a few hints about some phrases with *only*, and nothing like a uniform semantics analysis. For further discussion, cf. Keenan and Stavi [1986], Rooth [1984; 1985].

2.4.11 Exception DETs

This term is used by Keenan and Stavi for *DET*s like

all but three, all but at most five, all but finitely many, ...

As for interpretations, we have

all but three$_M AB \leftrightarrow |A - B| = 3$,
all but at most five$_M AB \Leftrightarrow |A - B| \leqq 5$,
all but finitely many$_M AB \Leftrightarrow A - B$ is finite.

The construction *all but Q* apparently obeys certain restrictions — we will return to these in 3.4. It can create ambiguities similar to the ones discussed for *only* in 2.4.10; cf.

(38) All but five liberal students voted for Smith.

[19] There is also the idea that five is unexpectedly few here. It would be possible to add $\mathbf{few}_M AB$ as a further condition.

There are also exception *DET*s with proper names and with definites:

every but John, no but John, every but John's, all but the liberal, ...

Some of these are discontinuous

(39) *Every* student *but John* voted for Smith,

(40) *Every* car *but John's* was stolen,

(41) *Every* book *but this* (one) was returned.

If we were to treat proper names as definites in the sense of 2.4.7, i.e. as providing suitable sets (in this case: the unit set of the denoted individual), we could interpret these on the uniform scheme

(42) **every but DEF**$_M AB \Leftrightarrow |X \cap A| = 1$ &
& **every**$_M A - XB$& **no**$_M A \cap XB$,

where, in (39), $X = \{\text{John}\}$, and, in (40), $X = P_{\text{John}}$; note that e.g. (39) says that John is a student, that he didn't vote for Smith, but that all other students voted for Smith. Note also that (42) gives conservative quantifiers.

2.4.12 Boolean combinations

First, *negation*, as in

not every, not all, not many, note more than five, not fewer than there, not more than half (of the), ...

The semantics of negated quantifiers is obvious,

(**not Q**)$_M \Leftrightarrow \neg$**Q**$_M AB$,

but *not* cannot stand in front of all *DET*s: e.g. *not some, not most, not at most five* are not well-formed. It is not clear that there is a semantic explanation for this. An interesting question, however, is whether the class of natural language quantifiers is *closed under negation*. For example, even though *not most* is not a *DET*, we can express the intended quantifier with another *DET*:

$$\begin{array}{lcl} \neg\ \textbf{most}_M AB & \Leftrightarrow & |A \cap B| \leqq |A - B| \\ & \Leftrightarrow & |A \cap B| \leqq 1/2|A| \text{ (on finite sets, of course)} \\ & \Leftrightarrow & \textbf{not more than half (of the)}_M AB \end{array}$$

Likewise, we have ¬(**at most five**) = **more than five**. But there are other cases which seem more doubtful, for example, the exception *DET*s: what *DET* would express the negation of *all but three* or *every but John*? We return to this question in 3.4.

As for conjunction and disjunction, we have

some but not all, some but not many, most but not all, at least five and at most ten, either exactly five or more than ten, neither less than five nor more than ten, John's but not Susan's, neither John's nor Susan's, both John's and Susan's, ...

Again the semantics is clear. It is tempting to claim that *any* two 1-place *DET*s can in principle be conjoined with *and* or *or* (another matter is that many such conjunctions and disjunction would be long and cumbersome, express trivial or otherwise 'strange' quantifiers, etc.) n-place *DET*s for $n > 1$ are discontinuous, which makes the claim less plausible in this case.[20] But the class of binary natural language quantifiers would, if the claim is correct, be closed under conjunction and disjunction.

Boolean operators can also be used to create n-place *DET*s for $n > 1$, e.g. the 2-place

every ... and, some ... or,

as in

(44) Every businessman and lawyer knows this,

(45) Some mother or father will react

Note that (43) is ambiguous. In general, there are two possible readings of sentences of the form QA *and/or* BC:

(45) $Q^1 A$ *and* $BC \leftrightarrow QA \cap BC$,
$Q^2 A$ *and* $BC \leftrightarrow QAC \,\&QBC$

(46) $Q^1 A$ *or* $BC \leftrightarrow QA \cup BC$,
$Q^2 A$ *or* $BC \leftrightarrow QAC \vee QBC$

In the one sense of (43) we have the ordinary *every* applied to the complex N *businessman and lawyer*, and in the other we have *every*^{2} applied to the two Ns *businessman* and *lawyer*. Of course, it is not absolutely necessary to use 2-place *DET*s here, since the interpretations are definable with 1-place *DET*s. For several arguments that 2-place *DET*s are in fact the natural choice, and for more examples, we refer to Keenan and Moss [1985].

We may note that

(47) *every*$^2 A$ *and* $BC \Leftrightarrow$ *every*$^1 A$ *or* BC,

(48) *some*$^1 A$ *or* $BC \Leftrightarrow$ *some*$^2 A$ *or* BC.

(47) explains why the second reading of (43) can also be expressed by

[20] We had a few examples of discontinuous 1-place *DET*s too, e.g. *every but John*, and here the claim is more dubious. But note that in all these cases, an alternative analysis was proposed, which eliminates the need for the *DET*s in question.

(49) Every businessman or lawyer knows this.

(48) explains why (44) isn't in fact ambiguous.

The same method as above can be used to create n-place *DET*s for all $n > 1$; cf.

(50) Every professor and assistant and secretary and student has a key.

This 4-place *DET* would be interpreted by a 5-ary quantifier similarly to (45) (the second reading seems to be preferred here, which again is manifested in the fact that *and* can be replaced by *or* in (50)).

3 QUANTIFIER CONSTRAINTS AND SEMANTIC UNIVERSALS

A natural way to approach the class of natural language quantifiers is to study the effect of linguistically motivated *constraints*, such as conservativity, on the class of all quantifiers. These constraints are related to *semantic universals*, i.e. general statements about semantic interpretation true for all natural languages. In this section we discuss some such constraints; a number of *possible* semantic universals will be noted along the way.

3.1 The Restriction to Monadic Quantifiers

In Section 2 we tacitly assumed that natural language quantifiers are monadic, i.e. of type $\langle 1, 1, \ldots, 1\rangle$. Is there some reason natural language should not employ non-monadic generalised quantifiers like those used in mathematical logic?

Towards an answer to this, recall first that generalised quantifiers are *second-order* properties or relations (cf. 1.2.1 and 1.4). Thus, *any* sentence which attributes, say, a (second-order) property to a (first-order) property can in principle be formalised as a quantified sentence. For example, consider

(1) Red is a colour.

Even in our extensional framework we *could* define a quantifier C of type $\langle 1\rangle$ by

$$C_M = \{X \subseteq M : X \text{ is the extension in } M \text{ of some colour}\}.$$

So C_M would contain the set of all blue things in M, the set of all red things in M, etc. Then (1) can be formalised as

(2) Cx *red*(x),

which is true in a model **M** iff the set which *red* denotes in **M** is (the extension of) a colour. This quantifier is monadic, but a similar story could be told for properties of binary relations, i.e. generalised quantifiers of type $\langle 2\rangle$.

But from our perspective, (2) is clearly an *unreasonable* formalisation of (1). It is useful to understand why. Compare (2) with

(3) $\exists x\ \mathit{red}(x)$,

which formalises

(4) Something is red.

There is a match in *logical form* between (3) and (4),[21] which is lacking between (1) and (2). Roughly, the difference is that *some* and *colour* are of completely different syntactic categories (*some* is an operator and *colour* is a predicate). In a natural language context, such matching appears to be essential. It is now always essential in mathematical contexts; cf. the quantifier **W**, where

$$Wxy Pxy$$

expresses that

P is a wellordering.

These remarks are really just another way of putting our basic idea that, in natural language, quantifier expressions are *DET*s. So the question is this: are there *DET*s denoting non-monadic quantifiers? Put differently, are there *DET*s whose corresponding quantifier symbols bind more than one variable in the succeeding formula(s)?

The following example was suggested by Hans Kamp:

(5) Most lovers will eventually hate each other.

This sentence makes good sense,[22] and, looking closely, one sees that it does not talk about the *set* of people who love and are loved by someone, but instead about *pairs*[23] of people who love each other: most such pairs will end up as pairs whose members hate each other. In other words, (5) is *not* equivalent to

(6) Most people who love and are loved by someone will eventually hate and be hated by everyone (or someone) they love.

This follows from the observation that one person may belong to different 'loving pairs'; using this it is easy to construct models where (5) and (6) (in either version) differ in truth value.[24]

[21] The match would be even better if we had used the binary **some** instead of the usual existential quantifier.

[22] Other similar sentences are harder to make sense of, for example,

Most schoolboys tease each other.

Is this about pairs of schoolboys, or does it mean that most schoolboys tease some other schoolboy, or most other schoolboys, ...? The problem seems to be that *schoolboy* denotes a set but *each other* indicates a relation.

[23] I take the pairs to be ordered, but this doesn't really matter.

[24] In other cases equivalence would obtain. Consider, for example,

Most twins like each other.

Since everyone is the twin of at most one other person, there are as many individual twins as there are ordered twin pairs, and thus the same proportion of 'liking' twin pairs as that of twins who like their other twin.

In the terminology of Section 1.4 we would formalise (5) as

(7) *most*$^{(2)}xy$(*love*$^*(x,y)$, *will eventually hate*$^*(x,y)$),

where $R^*(x,y)$ means $R(x,y) \wedge R(y,x)$ and

$$\mathbf{most}_M^{(2)} = \{\langle R_1, R_2\rangle : R_1, R_2 \subseteq M^2 \& \\ \& |R_1 \cap R_2| > |R_1 - R_2|\},$$

a generalised quantifier of type $\langle 2,2\rangle$.

Another suggestion to use quantification over pairs instead of individuals appears in Fenstad *et al.* [1987]. They consider sentences like

(8) Every boy who owns a dog kicks it.

There is a question as to the meaning of this, but the preferred reading appears to be that every boy who owns a dog kicks every dog he owns; in other words, using the binary *every* and *some*,

(9) *every* x(*boy*$(x)\wedge$ *some* y (*dog*(y), *owns*(x,y)),
every y(*dog*$(y)\wedge$ *owns* (x,y) *beats*(x,y)).

The traditional problem here has been to get (9) (or something equivalent to it) from a compositional analysis of (8); note that *it* refers back to *a dog*, but does not correspond to a bound variable in (9)! Fenstad *et al.* propose a way to do this; their analysis (whose details need not concern us here) leads, essentially, to the formalisation

(10) *every*$^{(2)}xy$(*boy*$(x)\wedge$ *dog*$(y)\wedge$ *owns*(x,y), *beats*(x,y))

where *every*$^{(2)}$ denotes the type $\langle 2,2\rangle$ generalised quantifier

(11) $\mathbf{every}_M^{(2)} = \{\langle R_1, R_2\rangle : R_1, R_2 \subseteq M^2 \& R_1 \subseteq R_2\}$.

Note that (10) and (9) are equivalent. (Note also, however that, as Johan van Benthem has pointed out, this analysis does not seem to work for all quantifiers: consider

(12) Most boys who own a dog kick it.

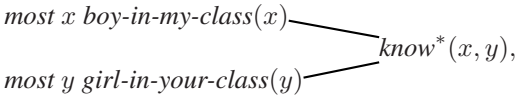
Here, the sentence obtained from (9) by replacing the first occurrence of *every* with *most* is *not* equivalent to the sentence obtained from (10) by replacing *every*$^{(2)}$ with *most*$^{(2)}$. Moreover, the former sentence appears to give the preferred reading.[25])

A third and final example that *could* be construed as quantification over pairs in natural language is branching quantification as discussed in Section 1.5. To take an example from Barwise [1979], consider

[25]Consider a situation with two boys, one of whom owns and kicks two dogs, the other owning, but not kicking, one dog. The formalisation with $\mathbf{most}^{(2)}$ would be true in this case, which seems counter-intuitive.

(13) Most boys in my class and most girls in your class know each other.

The preferred reading of this sentence (which has a *conjoined NP*) can be formalised as



(14)

where this is taken to mean that there is a subset X of the boys in my class, containing most of these boys, and a subset Y of the girls in your class, containing most of those girls, such that if $a \in X$ and $b \in Y$ then a knows$^* b$ (cf. Appendix A).

(14) involves branching of the ordinary monadic **most**. But, as noted in 1.5, it is *possible* to 'simulate' branching of two (or more) quantifiers by means of one generalised quantifier. That generalised quantifier will be non-monadic — in the present case, it has type $\langle 1, 1, 2\rangle$, since it relates two sets (the set of the boys and the set of the girls) and one binary relation (know*).

What can be concluded from these examples? Two things should be noted. The first is that the *logical power of expression increases* if the constructions in the examples are included. Consider the logic $L(\mathbf{most}^{(2)})$. It is easy to see that **most** is expressible in this logic, so $L(\mathbf{most} \leqq L(\mathbf{most}^{(2)})$. But the converse does not hold; the following result was pointed out by Per Lindström:

THEOREM 15. $L(\mathbf{most}) < L(\mathbf{most}^{(2)})$ *(*even on finite models*).*

Proof. [Cf. Section 1.7] Given a natural number d, choose two finite models $\mathbf{M}$= $\langle M, A_0, A_1, A_2\rangle$ and $\mathbf{M}' = \langle M', A'_0, A'_1, A'_2\rangle$ such that the $A_i(A'_i)$ are pairwise disjoint sets whose union is $M(M')$, and, if $|A_0| = k, |A_1| = m, |A_2| = n$, then $|A'_0| = k-1, |A'_1| = m, |A'_2| = n$, and

(a) $(k-1)m \leqq n < km$,

(b) $k < m < n$ and $k, m-k, n-m > 2d$.

Now, consider the sentence

$$\mathit{most}^{(2)}xy((P_0x \wedge P_1y) \vee (P_2x \wedge x = y), P_0x \wedge P_1y).$$

In **M**, this expresses that

$$km > n$$

(note that $P_2x \wedge x = y$ denotes $\{(a, a) \in M^2 : a \in A_2\}$, whose cardinal is n). Likewise, it expresses in $\mathbf{M}'$ that $(k-1)m > n$, so, by (a), it is true in **M** but false in $\mathbf{M}'$. On the other hand, using (b) and Theorem 10 it is easily seen that $\mathbf{M} \equiv_{d,\mathbf{most}} \mathbf{M}'$. Thus, since d was arbitrary, $L(\mathbf{most}^{(2)}) \not\leqq L(\mathbf{most})$. ■

The same holds for the branching of **most**. Let $L_b(\mathbf{most})$ be the logic which extends $L(\mathbf{most})$ by allowing formulas of the form (14), interpreted as indicated for that example. It can be shown that '$|A|$ is even' is expressible in $L_b(\mathbf{most})$. Thus, by (4) in Section 1.7, we get the

THEOREM 16. $L(\mathbf{most}) < L_b(\mathbf{most})$ *(even on finite models).*

The second observation to make, however, is that there are clear senses in which the non-monadic quantification considered here is *reducible* to monadic quantification. Thus, branching maybe seen as a linguistic construction on its own, making monadic quantifiers as *arguments*. And as for the first two examples, $\mathbf{most}^{(2)}$ is really just the old **most** applied to the new universe M^2:

$$\mathbf{most}^{(2)}_M = \mathbf{most}_{M^2},$$

and similarly for $\mathbf{every}^{(2)}$. Here we have *lifted* a relation on sets to a relation on binary relations. In general, any k-ary monadic quantifier $\mathbf{Q}$ can be lifted to any $n > 1$: define $\mathbf{Q}^{(n)}$, of type $\langle n, n, \ldots, n\rangle$ by letting, for all $R_1, \ldots, R_k \subseteq M^n$,

$$\langle R_1, \ldots, R_k\rangle \in \mathbf{Q}^{(n)}_M \Leftrightarrow \mathbf{Q}_{M^n} R_1, \ldots, R_k.$$

In view of the foregoing discussion we have a possible semantic universal of the form

($U1$) *Natural language quantifiers are either monadic or reducible to monadic quantifiers*,

where 'reducible' may be specified along the lines suggested above.

NB. This universal has been challenged recently, however, in [Keenan, 1987]. He considers sentences like

(15) Every boy read a different book

and shows that, although this may seem as simple *iteration* of two monadic quantifiers, the truth conditions for (15) cannot be so obtained, nor can they be obtained by branching or lifting monadic quantifiers. For further discussion of this matter, cf. also [van Benthem, 1987b]. In what follows, however, we will restrict attention to monadic quantifiers.

3.2 The Universe of Quantification

Recall the definition of conservativity for an $(n+1)$-ary quantifier $\mathbf{Q}$:

$$\textit{CONSERV}\ \mathbf{Q}_M A_1 \ldots A_n, B \Leftrightarrow \mathbf{Q}_M A_1 \ldots A_n, (A_1 \cup \ldots \cup A_n) \cap B$$

(for all M and all $A_1, \ldots, A_n, B \subseteq M$; we will usually omit this). We have put a comma before 'B' here to indicate that '$\mathbf{Q}_M A_1, \ldots A_n$' corresponds to the NP and 'B' to the VP. *CONSERV* says that the VP denotation can be restricted to (the union of) the N denotation(s). Another way to put this is

(*) *If B and C have the same intersections with all the A_i, then* $\mathbf{Q}_M A_1 \dots A_n, B \Leftrightarrow \mathbf{Q}_M A_1 \dots A_n, C$.

It is easily checked that *CONSERV* and (*) are equivalent conditions.

It is *almost* true that *CONSERV* restricts the universe of quantification to (the union of) the first (n) argument(s); cf. the discussion in 2.2.1. But not quite: the *DET* denotation may depend essentially on the universe M. The following condition, which we formulate for arbitrary n-ary quantifiers, expresses the requirement of 'universe-independence' for quantifiers ('*EXT*' for 'extension'):

EXT *If* $A_1, \dots, A_n \subseteq M \subseteq M'$
then $\mathbf{Q}_M A_1 \dots A_n \Leftrightarrow \mathbf{Q}_{M'} A_1 \dots A_n$.

This has nothing to do with *CONSERV*; rather, it is a strengthening of the postulate, discussed in 2.1.3; that quantifier expressions are *constants*. For example, *EXT* excludes a quantifier which is $\mathbf{all}_M$ when M has fewer than 10 elements and $\mathbf{some}_M$ otherwise. But *together* with *CONSERV, EXT* gives the exact sense in which *DET*s can be said to restrict the universe of quantification:

UNIV $\mathbf{Q}_M A_1 \dots A_n, B \Leftrightarrow$
$\mathbf{Q}_{A_1 \cup \dots \cup A_n} A_1 \dots A_n, (A_1 \cup \dots \cup A_n) \cap B$.

It is an easy exercise to show

PROPOSITION 17. *UNIV is equivalent to CONSERV + EXT.*

Some further discussion of universe-restriction can be found in Westerståhl [1985a; 1983].

CONSERV and *EXT* are related to the logician's notion of *relativisation* (Sections 1.4 and 1.6). Let us first note

PROPOSITION 18. *If* $\mathbf{Q}^i$ *satisfies EXT for* $i \in I$, *then* $L(\mathbf{Q}^i)_{i \in I}$ *relativises.*

Proof. Since *EXT* implies that

$$(Q^i)^r x(Px, P_1 x, \dots, P_n x) \leftrightarrow$$
$$\leftrightarrow Q^i x(Px \wedge P_1 x, \dots, Px \wedge P_n x)$$

is valid. ■

If in addition *CONSERV* holds we can say more: the *binary* quantifiers satisfying *CONSERV* and *EXT* are precisely the relativized ones. Moreover, the *sentences* (in any logic) with two unary predicate symbols which satisfy *CONSERV* and *EXT* (in the obvious sense) are precisely the ones equivalent to the relativised sentences. This is the content of the next result:

THEOREM 19.

(a) A binary quantifier $\mathbf{Q}$ *satisfies CONSERV and EXT iff* $\mathbf{Q} = (\mathbf{Q}')^r$, *for some unary* $\mathbf{Q}'$.

(b) A sentence $\phi(P_1, P_2)$ with two unary predicate symbols in a logic L satisfies CONSERV and EXT iff it is equivalent to $\psi^{(P_1)}$, for some L-sentence ψ.

Proof. We prove (b); (a) then follows (it is also easily proved directly). Recall the basic property of relativised sentences from 1.6, in this case, with $\mathbf{M}= \langle M, A, B\rangle$,

$$(REL) \quad \langle M, A, B\rangle \vDash \psi^{(P_1)} \Leftrightarrow \langle A, A \cap B\rangle \vDash \psi.$$

From this it is immediate that $\psi^{(P_1)}$ satisfies *CONSERV* and *EXT*. Conversely, if $\phi(P_1, P_2)$ satisfies *CONSERV* and *EXT*, let $\psi = \phi(x = x, P_2)$.

Then

$$\begin{array}{rcll} \langle M, A, B\rangle \vDash \psi^{(P_1)} & \Leftrightarrow & \langle A, A \cap B\rangle \vDash \psi & (\textit{REL}) \\ & \Leftrightarrow & \langle A, A, A \cap B\rangle \vDash \phi(P_1, P_2) & (\text{by def. of } \psi) \\ & \Leftrightarrow & \langle M, A, B\rangle \vDash \phi(P_1, P_2) & (\text{by } \textit{UNIV}). \end{array}$$

■

The interest of (b) is that it relates a semantic notion (*CONSERV* and *EXT*) to a syntactic property of sentences — a typical sort of logical result.

Notice that, for *unary* quantifiers, *CONSERV* makes no sense, and *EXT*, although it can be formulated, is *not true* for e.g. the standard universal quantifier $\forall$. This is another aspect of the advantage of binary quantifiers. Any unary quantifier can be replaced by a binary one (its relativisation) which does (at least) the same work and has the additional property of restriction the universe of quantification to the first argument. As Theorem 19 shows, this moves give us *all* the binary quantifiers with that property, in particular, it gives us all the binary natural language quantifiers (provided $(U2)$ and $(U3)$ below hold).

For n-ary quantifiers with $n > 1$, it is also possible to secure *CONSERV* and *EXT* by raising the number of arguments, though not quite as simply as when $n = 1$. The next proposition surveys the possibilities.

PROPOSITION 20. *Let $Q\mathbf{Q}$ be an n-ary quantifier. then*

(i) there is an $(n+1)$-ary quantifier $\mathbf{Q}'$ satisfying CONSERV such that $\mathbf{Q}_M A_1 \ldots A_n \Leftrightarrow \mathbf{Q}'_M A_1 \ldots A_n, M$;

(ii) there is an $(n+2)$-ary quantifier $\mathbf{Q}''$ satisfying EXT such that $\mathbf{Q}_M A_1 \ldots A_n \Leftrightarrow \mathbf{Q}''_M A_1 \ldots A_n, M$;

(iii) there is an $(n+1)$-ary quantifier $\mathbf{Q}^+$ satisfying both CONSERV and EXT such that $\mathbf{Q}_M A_1 \ldots A_n, \Leftrightarrow \mathbf{Q}^+_M A_1 \ldots A_n M, M$.

Proof.

(i) Define $\mathbf{Q}'_M A_1 \ldots A_n, B \Leftrightarrow \mathbf{Q}_M A_1 \ldots A_n \cap B \ldots A_n \cap B$. The verification of *CONSERV* is immediate.

(ii) Let $\mathbf{Q}''_M A_1 \dots A_n, B \Leftrightarrow \mathbf{Q}_B A_1 \cap B \dots A_n \cap B$; again *EXT* is immediate.

(iii) Define $\mathbf{Q}''$ as in (ii), and then form $\mathbf{Q}^+$ from $\mathbf{Q}''$ as in (i); the result follows from (i) and (ii).

■

For the record, we formulate the semantic universals corresponding to *CONSERV* and *EXT*:

($U2$) Natural language quantifiers are conservative.

($U3$) Natural language quantifiers satisfy *EXT*.

We saw in 2.4 that the few apparent exceptions to ($U2$) could be accounted for by reasonable methodological decisions (2.4.3, 2.4.9–10). As for ($U3$), the only exceptions found were certain interpretations of context-dependent *DET*s like *many*. For example, if

$$\mathbf{Q}_M AB \Leftrightarrow |A \cap B| \geqq 1/3|M|,$$

$\mathbf{Q}$ violates *EXT*. Again, it is mainly a methodological question whether one wants to allow this kind of context-dependence or not.

3.3 Quantity

The condition *ISOM*, repeated below, was formulated for generalised quantifiers of any type $\langle k_1, \dots, k_n \rangle$:

ISOM *If f is a bijection from M to M'*
then $\mathbf{Q}_M R_1 \dots R_n \Leftrightarrow \mathbf{Q}_{M'} f[R_1] \dots f[R_n]$.

The idea is that $\mathbf{Q}$ does not distinguish between different elements of the universe, or even across two universes. This requirement, which is a version of what is sometimes called *topic-neutrality*, can be formulated for arbitrary syntactic categories (cf. [van Benthem, 1983b]). It is a general requirement of *logical constants*.

For monadic quantifiers, *ISOM* has a particularly conspicuous formulation. Roughly, it says that quantifiers deal only with *quantities*. The latter assertion can be made precise with the terminology from Section 1.7 as follows:

QUANT *If* $\mathbf{M} = \langle M, A_0, \dots, A_{k-1} \rangle$, $\mathbf{M}' = \langle M', A'_0, \dots, a'_{k-1} \rangle$, *and* $|P_s^{\mathbf{M}}| = |P_s^{\mathbf{M}'}|$ *for all* $s \in 2^k$, *then* $\mathbf{Q}_M A_0 \dots A_{k-1} \Leftrightarrow \mathbf{Q}_{M'} A'_0 \dots A'_{k-1}$.

This means that the truth value of $\mathbf{Q}_M A_0 \dots A_{k-1}$ depends only on w^k quantities, namely, the number of elements in the partition sets.

A bijection from M to M' splits into bijections of the respective partition sets, and, conversely, bijections between these sets can be joined to one from M to M'. Thus we have that

PROPOSITION 21. *ISOM and QUANT are equivalent (for a monadic* **Q***).*

If we consider only one universe M in *ISOM* (letting $M' = M$), and thus *permutations* on M, we get a slightly weaker version, called *PERM*.[26] From a local perspective on quantifiers (2.1.4), *PERM* is the natural notion. Our global condition *EXT*, however, says that the choice of universe is unimportant. Indeed, it is straightforward to prove

PROPOSITION 22. *Under EXT, ISOM and PERM are equivalent.*

All the simplex *DET*s from 2.4.1–3 denote quantitative quantifiers. To see this, it is sufficient to check that the defining conditions can be expressed as conditions on the cardinalities of the relevant sets. For example, $\mathbf{all}_M AB \Leftrightarrow |A - B| = 0, \mathbf{some}_M AB \Leftrightarrow |A \cap B| \neq 0, \mathbf{most}_M AB \Leftrightarrow |A \cap B| > |A - B|, \mathbf{both}_M AB \Leftrightarrow |A - B| = 0 \& |A \cap B| = 2, \mathbf{many}^2_M AB \Leftrightarrow |A \cap B| \geqq k(|A \cap B| + |A - B|)$, etc.

As for complex *DET*s, there are just a few of the constructions in 2.4.6–12 which yield non-quantitative quantifiers. One example is *DET*s with fixed adjective phrases, or similar expressions, such as *more male than female, some red, only liberal*. We saw, however, that sentences with such expressions can also be interpreted using only quantitative quantifiers (2.4.9–10). Another major example are the possessives, either simple ones such as *John's*, or complex constructions with possessives. The quantifier **John's** from 2.4.6 violates ISOM since the ownership relation need not be preserved under permutations of the objects in the universe. For example, John may own two white shirts but no red tie, even though it is possible to permute the shirts and the ties, and the white things and the red things in a one-one fashion. Then

John's shirts are white

is true, but not

John's ties are red,

as *ISOM* would require.

In 2.4.7, we mentioned an alternative analysis of definites, and thus in particular of possessives. Under this analysis, one can dispense with quantifiers denoted by simple possessives, also in various complex constructions. Quantitative quantifiers would suffice, it seems, for all of these constructions (the same holds for *every but John* (2.4.11), another counter-instance to *ISOM*). It would then be possible to propose the following rather appealing universal:

($U4$) *Natural language quantifiers are quantitative.*

If one does not want to take this methodological step, on the other hand, one will settle for the more modest

($U4'$) *Simple natural language quantifiers are quantitative.*

[26] To get a 'quantity version' of *PERM*, let $M' = M$ in *QUANT*.

3.4 Logical Quantifiers, Negations and Duals

Whichever version of the last universal one prefers, the following class of quantifiers is a natural object of study:

DEFINITION 23. If n-ary quantifier $(n > 1)$ is *logical* then it satisfies *CONSERV, EXT* and *QUANT*.

The terminology is meant to suggest that these three requirements are *necessary* for logicality; further conditions will be discussed in 4.4.

For *binary* quantifiers, logicality means that the truth value of $\mathbf{A}_M AB$ depends only on the two numbers $|A - B|$ and $|A \cap B|$:

PROPOSITION 24. *A binary quantifier* $\mathbf{Q}$ *is logical iff, for all* M, M' *and all* $A, B \subseteq M$ *and* $A', B' \subseteq M', |A - B| = |A' - B'|$ *and* $|A \cap B| = |A' \cap B'|$ *implies that* $\mathbf{Q}_M AB \Leftrightarrow \mathbf{Q}_{M'} A'B'$.

Proof. If $\mathbf{Q}$ is logical and $|A - B| = |A' - B'|$ and $|A \cap B| = |A' \cap B'|$, then, by *QUANT*, $\mathbf{Q}_A AA \cap B \Leftrightarrow \mathbf{Q}_{A'} A'A' \cap B'$, and so, by *UNIV* (Proposition 17), $\mathbf{Q}_M AB \Leftrightarrow \mathbf{Q}_{M'} A'B'$. Conversely, if the right-hand side of the equivalence holds, *QUANT* is immediate. Take M and $A, B \subseteq M$ and let $M' = A' = A$ and $B' = A \cap B$. Thus, $\mathbf{Q}_M AB \Leftrightarrow \mathbf{Q}_A AA \cap B$, i.e. *UNIV* holds. ■

This means that a logical binary relation between *sets* can be replaced by a binary relation between *cardinal numbers*; we exploit this in 4.2. Proposition 24 can be generalised to n-ary logical quantifiers: *QUANT* transforms an n-ary $\mathbf{Q}$ to a relation between 2^n cardinal numbers, and *CONSERV* + *EXT* eliminate the dependence of *two* of these.

The class of logical quantifiers has some nice closure properties. It is straightforward to verify that if $\mathbf{Q}_1$ and $\mathbf{Q}_2$ are *CONSERV and EXT* (*QUANT*), then so are $\mathbf{A}_1 \wedge \mathbf{Q}_2, \mathbf{Q}_1 \vee \mathbf{Q}_2$, and $\neg\mathbf{Q}_1$. Thus,

PROPOSITION 25. *For each* $n > 1$*, the class of* n*-ary logical quantifiers is closed under the usual Boolean operations.*

In a natural language context, there are also *inner* Boolean operations. We noted in 2.4.12 that from a binary $\mathbf{Q}$ one can construct two $(n + 1)$-ary inner conjunctions:

$$\mathbf{Q}_M^{\wedge 1} A_1 \ldots A_n, B \Leftrightarrow \mathbf{Q}_M A_1 \cap \ldots \cap A_n B,$$
$$\mathbf{Q}_M^{\wedge 2} A_1 \ldots A_n, B \Leftrightarrow \mathbf{Q}_M A_1 B \& \ldots \& \mathbf{Q}_M A_n B.$$

Inner disjunctions $\mathbf{Q}^{\vee 1}$ and $\mathbf{Q}^{\vee 2}$ are defined similarly. As for negation, we make the

DEFINITION 26. If $\mathbf{Q}$ is $(n + 1)$-ary, the *inner negation* of $\mathbf{Q}$ is the quantifier $\mathbf{Q}\neg$, defined by

$$(\mathbf{Q}\neg)_M A_1 \ldots A_n B \Leftrightarrow \mathbf{Q}_M N A_1 \ldots A_n, M - B.$$

Also, the *dual* of $\mathbf{Q}$ $\check{\mathbf{Q}}$, is the quantifier $\neg(\mathbf{Q}\neg)(= (\neg\mathbf{Q})\neg)$.

Outer and inner negation correspond to sentence negation and *VP* negation, respectively; cf.

Not many boys are lazy,
Many boys are not lazy,

with the respective truth conditions

$(\neg\mathbf{many})_M\|boy\|\ \|lazy\|$,
$(\mathbf{many}\neg)_M\|boy\|\ \|lazy\|$.

PROPOSITION 27. *The class of logical quantifiers is closed under inner conjunctions and disjunctions (both kinds), and inner negation (hence also duals).*

Proof. This is again a routine check; let us take one case and verify that $\mathbf{Q}\neg$ satisfies *EXT* if $\mathbf{Q}$ satisfies *CONSERV* and *EXT*. Suppose $A_1, \ldots, A_n, B \subseteq M \subseteq M'$. Then

$$\begin{aligned}
(\mathbf{Q}\neg)_M A_1 \ldots A_n, B &\Leftrightarrow \mathbf{Q}_M A_1 \ldots A_n, M - B \\
&\Leftrightarrow \mathbf{Q}_M A_1 \ldots A_n, (A_1 \cup \ldots \cup A_n) - B && (\textit{CONSERV}) \\
&\Leftrightarrow \mathbf{Q}_{M'} A_1 \ldots A_n, (A_1 \cup \ldots \cup A_n) - B && (\textit{EXT}) \\
&\Leftrightarrow \mathbf{Q}_{M'} A_1 \ldots A_n, M' - B && (\textit{CONSERV}) \\
&\Leftrightarrow (\mathbf{Q}\neg)_{M'} A_1 \ldots A_n, B.
\end{aligned}$$

■

It should be noted that other inner negations than *VP* negation do *not* preserve logicality. For example, if we define,for a binary $\mathbf{Q}$,

$$\mathbf{Q}^*_M AB \Leftrightarrow \mathbf{Q}_M M - AB,$$

then *CONSERV* will not be preserved.

The following propositions list some de Morgan-like laws for inner Boolean operations on quantifiers:

PROPOSITION 28.

(a) $(\neg\mathbf{Q})^{\wedge 1} = \neg(\mathbf{Q}^{\wedge 1}), (\neg\mathbf{Q})^{\wedge 2} = \neg(\mathbf{Q}^{\vee 2})$,

(b) $(\neg\mathbf{Q})^{\vee 1} = \neg(\mathbf{Q}^{\vee 1}), (\neg\mathbf{Q})^{\vee 2} = \neg(\mathbf{Q}^{\wedge 2})$,

(c) $(\mathbf{Q}_1 \wedge \mathbf{Q}_2)\neg = \mathbf{Q}_1\neg \wedge \mathbf{Q}_2\neg$,

(d) $(\mathbf{Q}_1 \vee \mathbf{Q}_2)\neg = \mathbf{Q}_1\neg \vee \mathbf{Q}_2\neg$,

(e) $(\mathbf{Q}^{\wedge i})\neg = (\mathbf{Q}\neg)^{\wedge i} (i = 1, 2)$,

(f) $(\mathbf{Q}^{\vee i})\neg = (\mathbf{Q}\neg)^{\vee i} (i = 1, 2)$,

In 2.4.12 we considered the suggestion that the class of binary natural language quantifiers is closed under (outer) conjunction and disjunction, i.e. that the following universal holds:

($U5$) *If* $\mathbf{Q}_1$ *and* $\mathbf{Q}_2$ *are binary natural language quantifiers then so are* $\mathbf{Q}_1 \wedge \mathbf{Q}_2$ *and* $\mathbf{Q}_1 \vee \mathbf{Q}_2$.

The case of negation was more doubtful. In the table opposite, some examples of *DET*s for negations and duals in English are given. '-' means that it seems hard to find a *DET*, simplex or complex, denoting the negation or dual in question. Of course these quantifiers are always expressible by some suitable paraphrase, but the question here is whether there are *determiners* denoting them.

This table suggests certain questions. When is the (inner or outer) negation of a simple quantifier again simple? Barwise and Cooper have several proposals here, e.g. that the negations of the cardinal quantifiers **at least n** and **exactly n** re never simple, and that if a language has a pair of simple duals, that pair consists of **every** and **some**; cf. also 3.6.

Here we shall look a bit closer at the '-' signs for the binary quantifiers in the table. Note that if these signs are correct, the class of binary natural language quantifiers is not closed under inner or outer negation. Discussing this question will give us an occasion to look at some typical issues, and to introduce a few useful notions. The purpose, as usual, is to illustrate problems and ideas, rather than making definite empirical claims.

Table 1.

Q	**¬Q**	**Q¬**	**˘Q**
some	*no*	*not every*	*every*
every	*not every*	*no*	*some*
no	*some*	*every*	*not every*
most	*at most half*	*less than half*	*at least half*
many	*few*	-	*all but a few*
infinitely many	*at most finitely many*	-	*all but finitely many*
(at least) n	*less than n*	-	*all but less than n*
at most n	*more than n*	*all but at most n*	-
(exactly) n	*not exactly n*	*all but n*	-
more . . . than	*at most as many . . . as*	-	-
fewer . . . than	*at least as many . . . as*	-	-

Note first that part of what Table 1 claims is that certain expressions of the form *all but Q* are anomalous. Thus, while *all but five, all but at most five, all but finitely many* are fine, *all but at least five, all but not exactly five, all but (infinitely) many* are not. It might be claimed that the anomaly in the latter cases is pragmatic rather than semantic. I will not argue about this directly, but instead try to see if there

are in fact significant semantic differences between the normal and the anomalous cases.

Exception *DET*s of the form *all but Q* (cf 2.4.11) are interpreted on the scheme

(1) **all but Q** $= \mathbf{Q}\neg$.

When is $\mathbf{Q}\neg$ a natural language quantifier? Before trying to give some answers to this, we need to introduce a new concept.

DEFINITION 29. A binary quantifier $\mathbf{Q}$ is *VP-positive* (*VP-negative*) if, for all M, M' and all $A, B \subseteq M, A'B' \subseteq M'$ such that $A \cap B = A' \cap B'(A - B = A' - B'), \mathbf{Q}_M AB \leftrightarrow \mathbf{Q}_{M'} A; B'$.[27]

As the terminology indicates, *VP*-positivity means that $\mathbf{Q}$ amounts solely to a condition on the *VP* denotation (intersected with the N denotation, since we assume *CONSERV*), whereas a *VP*-negative quantifier reduces to a condition on the *complement* of the *VP* denotation. For example, **some, no, many**,[28] **few, infinitely many, at least n, at most n, exactly n** are *VP*-positive, whereas **every, not every, all but n, all but at most n** are *VP*-negative. **most, at least half**, and other 'proportional' quantifiers are neither *VP*-positive nor *VP*-negative, and the same holds for the interpretations of the definites (because of the *number condition* on the N denotation; cf. 2.4.6–7).

For a *conservative* $\mathbf{Q}$, *VP*-positivity (-negativity) is related to inner and outer negation as follows:

(2) $\mathbf{Q}$ is *VP*-positive (-negative) $\Leftrightarrow \neg\mathbf{Q}$ is *VP*-positive (-negative)
$\Leftrightarrow \mathbf{Q}\neg$ is *VP*-negative (-positive).

The next result, essentially due to Barwise and Cooper shows that *VP*-positivity is in fact a simple relational property of quantifiers. A binary quantifier is *symmetric* if it is symmetric as a relation, i.e. iff for all M and all $A, B \subseteq M$,

$$\mathbf{Q}_M AB \Rightarrow \mathbf{Q}_M BA.$$

PROPOSITION 30. *If* $\mathbf{Q}$ *satisfies CONSERV and EXT the following are equivalent:*

(a) $\mathbf{Q}$ *is* VP*-positive.*

(b) $\mathbf{Q}$ *is symmetric.*

(c) $\mathbf{Q}_M AB \leftrightarrow \mathbf{Q}_M A \cap BA \cap B$ *(for all* M *and all* $A, B \subseteq M$*).*

[27] VP-positivity is related to the notions of *existential* and *cardinal* quantifiers in [Keenan and Stavi, 1986]. In fact, under *CONSERV*, VP-positivity is equivalent to existentiality, and cardinality is equivalent to VP-positivity $+$*QUANT*.

[28] This is for **many**^{1}(2.4.3); **many**^{2} is neither VP-positive nor VP-negative.

Proof. (a) $\Rightarrow$ (b): Suppose $\mathbf{Q}_M AB$. Let $A' = B$ and $B' = A$. Thus $A \cap B = A' \cap B'$, so, by *VP*-positivity, $\mathbf{Q}_M AB'$, i.e. $\mathbf{Q}_M BA$.

(b)$\Rightarrow$ (c): Suppose $\mathbf{Q}$ is symmetric. Then $\mathbf{Q}_M AB \Leftrightarrow \mathbf{Q}_M AA \cap B$ (*CONSERV*) $\Leftrightarrow \mathbf{Q}_M A \cap BA$ (symmetry) $\Leftrightarrow \mathbf{Q}_M A \cap BA \cap B$ (*CONSERV*).

(c) $\Rightarrow$ (a): If (c) holds and $A \cap B = A' \cap B'$, where $A, B \subseteq M$ and $A', B' \subseteq M'$, then $\mathbf{Q}_M AB \Leftrightarrow \mathbf{Q}_M A \cap BA \cap B \Leftrightarrow \mathbf{Q}_M A' \cap B' A' \cap B' \Leftrightarrow \mathbf{Q}_{M'} A' \cap B' A' \cap B$; (by *EXT*) $\Leftrightarrow \mathbf{A}_{M'} A' B'$ (by (c)). ■

The following corollary is easy using (2):

COROLLARY 31. *Under CONSERV and EXT the following are equivalent:*

(a) $\mathbf{Q}$ *is VP-negative.*

(b) $\mathbf{Q}^{\neg}$ *is symmetric.*

(c) $\mathbf{Q}_M AB \Leftrightarrow \mathbf{Q}_M A - B\varnothing$.

From our list of English *DET*s in 2.4, it appears much easier to find *VP*-positive quantifiers than *VP*-negative ones. Moreover, it seems that for each *DET* giving a condition on the complement of the *VP* denotation, there is another *DET* giving the *same* condition on the *VP* denotation itself. For example, if the first *DET* is of the form *all but q*, the corresponding positive condition is given by Q, and if the first *DET* is *every* or *not every*, the second is *no* or *some*, respectively. This lets us propose the following universal:

($U6$) *If* $\mathbf{Q}$ *is a VP-negative natural language quantifier, then* $\mathbf{Q}^{\neg}$ *is also a natural language quantifier.*

A related observation is that when Q denotes a *VP*-negative quantifier, the form *all but Q* is not allowed: *all but every, all but not every, all but all but five*, etc. are ruled out. The reason, one imagines, is that this would be a very cumbersome way of expressing a 'double *VP* negation', which in any case is equivalent to the more easily expressed positive condition.

($U6$) gives one (partial) answer to our question about when $\mathbf{Q}^{\neg}$ is a natural language quantifier. But, to return to Table 1, the most interesting case concerns *VP*-positive quantifiers: all the '-' signs (for binary quantifiers) are examples of failure of $\mathbf{Q}^{\neg}$ to be a natural language quantifier for *VP*-positive $\mathbf{Q}$. What, then, is wrong with a *DET* such as *all but at least five*?

Here is one suggestion: sentences of the form *all but Q A B* imply the *existence* of *A*s that are B (in contrast with *all A B*). More precisely, let us say that a quantifier $\mathbf{Q}$ has *existential import*, if

(3) for sufficiently large A (and M), $\mathbf{Q}_M AB \Rightarrow \mathbf{some}_m AB$.

(3) holds for **all but five, all but at most five, all but finitely many**, etc., but fails for (**at least five**)$\neg$, (**not exactly five**)$\neg$, (**infinitely many**)$\neg$, etc. E.g.

$$(\textbf{at least five})^{\neg}{}_M AB \Leftrightarrow |A - B| \geqq 5,$$

so for each A with at least five elements, we have (**at least five**)$\neg_M A\varnothing$ but not $\textbf{some}_M A\varnothing$. Note that the qualification 'for sufficiently large A' is necessary: $\textbf{all but at most five}_M AB$ implies $\textbf{some}_M AB$ only when $|A| > 5$, and $\textbf{all but finitely many}_M AB$ implies $\textbf{some}_M AB$ only when A is infinite.

What condition on $\mathbf{Q}$ corresponds to the fact that $\mathbf{Q}\neg$ has existential import/ For *VP*-positive quantifiers, the answer is as follows. Call $\mathbf{Q}$ *bounded*, if

(4) there is an n such that for all M and all $A, B \subseteq M$, $\mathbf{Q}_M AB \Rightarrow |A \cap B| \leqq n$.

PROPOSITION 32. *Suppose* $\mathbf{Q}$ *is VP-positive and satisfies CONSERV and EXT. Then* $\mathbf{Q}\neg$ *has existential import iff* $\mathbf{Q}$ *is bounded.*

Proof. If $\mathbf{Q}$ is bounded by n, then $|A| > n \& \mathbf{Q}\neg_M AB \Rightarrow |A| < n \& |A - B| \leqq n \Rightarrow A \cap B \neq \varnothing$, so (3) holds for $\mathbf{Q}\neg$. On the other hand, if $\mathbf{Q}$ is not bounded, it follows from proposition 30 that there are arbitrarily large A (and M) such that $\mathbf{Q}_M AA$. But this means that $\mathbf{Q}\neg_M A\varnothing$, so (6) fails for $\mathbf{Q}\neg$. ■

From these observations it is tempting to suggest the universal: for *VP*-positive $\mathbf{Q}$, $\mathbf{Q}\neg$ is a natural language quantifier only if $\mathbf{Q}$ is bounded. But this would be premature. The universal concerns arbitrary quantifiers $\mathbf{Q}\neg$, whereas the above discussion concerned the interpretations of *DET*s of the form *all but* Q'. In fact, there is a simple counter example to this universal: **some** is *VP*-positive, $\textbf{some}\neg =$ **not every** is a natural language quantifier, but **some** is not bounded!

Of course we cannot require in the universal that $\mathbf{Q}\neg$ be the interpretation of a *DT all but* Q'; that would make $\mathbf{Q}\neg$ *trivially* a natural language quantifier! But all is not lost: it seems that if we require $\mathbf{Q}$ to be *non-simple*, the universal holds; possibly, the simple **some** was the *only* counter-example.

What about the converse statement, i.e. if $\mathbf{Q}$ is bounded, does it follow that $\mathbf{Q}\neg$ is a natural language quantifier? Here we can say something more definite:

PROPOSITION 33. *If* $\mathbf{Q}$ *is logical, VP-positive, and bounded, then* $\mathbf{Q}$ *is a finite disjunction of quantifies of the form* **exactly n**.

(The proof is best postponed until Section 4.2.) Thus if $\mathbf{Q}$ is as in this proposition, $\mathbf{Q}$ is clearly a natural language quantifier, and so is $\mathbf{Q}\neg$, which by Proposition 28 is a finite disjunction of quantifiers of the form **all but n**.

Some of the last observations are collected in the following tentative universal:

($U7$) *If* $\mathbf{Q}$ *is a VP-positive, non-simple, logical quantifier,then* $\mathbf{Q}\neg$ *is a natural language quantifier iff* $\mathbf{Q}$ *is bounded.*

This universal, then, would be an explanation of the empty spaces (for the binary quantifiers) in Table 1.

3.5 Non-Triviality

Call an n-ary quantifier $\mathbf{Q}$ *trivial on M*, if $\mathbf{Q}_M$ is either the empty or the universal n-ary relation on $P(M)$. Consider the condition

NONTRIV $\mathbf{Q}$ *is non-trivial on some universe.*

Quantifiers violating *NONTRIV* are not very interesting: either any sentence beginning with a *DET* denoting such a quantifier (satisfying *EXT*) is true in each model, or any such sentence is false in each model. Nevertheless, natural language permits the construction of such *DET*s, for example, *at least zero, fewer than zero, at least ten and at most nine, more than infinitely many*, as pointed out in [Keenan and Stavi, 1986]. But the following universal seems true:

$(U8)$ *Simple natural language quantifiers satisfy NONTRIV.*

Note that the *NONTRIV* quantifiers are *not* closed under Boolean operations: for any $\mathbf{Q}$, the quantifier $\mathbf{Q} \vee \neg\mathbf{Q}$ is trivial on every universe.

NONTRIV requires a very modest amount of 'activity' of $\mathbf{Q}$; a stronger variant is

ACT $\mathbf{Q}$ *is non-trivial on each universe.*

ACT holds for many natural language quantifiers, but there are exceptions even among the simple ones, e.g. **both, two, three, four**, ... (if M has less than 4 elements $\mathbf{four}_M AB$ is always false).

van Benthem [1984a] considers an even stronger requirement of activity, called 'variety', for binary quantifiers. Here is a generalisation to $(n+1)$-ary quantifiers:

VAR *For all M and all $A_1, \ldots, A_n \subseteq M$ such that $A_1 \cap \ldots \cap A_n \neq \varnothing$, there are $B_1, B_2 \subseteq M$ such that $\mathbf{Q}_M A_1, \ldots, A_n, B_1$ and $\neg\mathbf{Q}_M A_1, \ldots, A_n, B_2$.*

In the binary case, we could say that *VAR* transfers the requirement of activity to each non-empty first argument. For quantifiers satisfying *CONSERV* and *EXT*, this seems a reasonable strengthening of *ACT*.

Clearly,

$$\textit{VAR} \Rightarrow \textit{ACT} \Rightarrow \textit{NONTRIV};$$

the implications cannot be reversed: an example of a (logical) quantifiers satisfying *ACT* but not *VAR* is

$$\mathbf{Q}_M AB \Leftrightarrow |A| = 1.$$

Note that this does not seem to be a natural language quantifier. In fact, inspection of the *DET*s in 2.4 shows that the *ACT* ones — e;g; **some, no, all, not all, most, more ... than, fewer ... than, every ... and/or, some ... and/or** (both interpretations) — also satisfy *VAR*. So one may propose

$(U9)$ *Natural language quantifiers satisfying ACT also satisfy VAR.*

3.6 Monotonicity

The monotonicity behaviour of a quantifier **A** concerns the preservation of other truth value of $\mathbf{Q}_M A_1 \ldots, A_n$ when the arguments are decreased or increased. For simplicity, we shall only consider *binary* quantifiers here, although many of the definitions and results below can easily be extended to $(n+1)$-ary quantifiers.

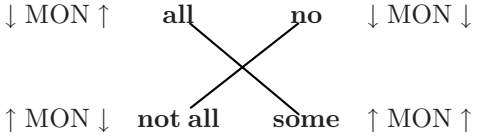
DEFINITION 34. A binary quantifier $\mathbf{Q}$ is

$$
\begin{array}{l}
MON\uparrow, \text{ if } \mathbf{Q}_M AB \& B \subseteq B' \Rightarrow \mathbf{Q}_M AB', \\
MON\downarrow, \text{ if } \mathbf{Q}_M AB \& B' \subseteq B \Rightarrow \mathbf{Q}_M AB', \\
\uparrow MON, \text{ if } \mathbf{Q}_M AB \& A \subseteq A' \Rightarrow \mathbf{Q}_M A'B, \\
\downarrow MON, \text{ if } \mathbf{Q}_M AB \& A' \subseteq A \Rightarrow \mathbf{Q}_M A'B.
\end{array}
$$

Also, $\mathbf{Q}$ is *RIGHT MON* (*LEFT MON*) if it is *MON*$\uparrow$ or *MON*$\downarrow$ ($\uparrow$*MON* or $\downarrow$*MON*), and $\mathbf{Q}$ is $\uparrow$*MON*$\uparrow$ if it is both *MON*$\uparrow$ and $\uparrow$*MON*; similarly for $\uparrow$*MON*$\downarrow$, $\downarrow$*MON*$\uparrow$, and $\downarrow$*MON*$\downarrow$.

Barwise and Cooper call *RIGHT MON* monotonicity, $\uparrow$*MON persistence* and $\downarrow$*MON anti-persistence*.

Many natural language quantifiers have simple monotonicity properties. The four types of *double monotonicity* are exemplified by the square of the opposition:



Other doubly monotone quantifiers are **at least n, infinitely many**, which are $\uparrow$*MON*$\uparrow$, and **at most n, at most finitely many, only liberal** (cf 2.4.10), which are $\downarrow$*MON*$\downarrow$. **most** is *MON*$\uparrow$ but not *LEFT MON*, as is easily seen, and the same holds for simple definites like **the** and **John's** (as defined in 2.4.6). Of the interpretations of *any* from 2.4.3, **many**^{1} is $\uparrow$*MON*$\uparrow$, **many**^{2} is *MON*$\uparrow$ but not *LEFT MON*, and **many**^{3} is neither *LEFT* nor *RIGHT MON*. Other examples of neither *LEFT* nor *RIGHT MON* quantifiers are **exactly n, all but n, between five and ten**.

The monotonicity behaviour of $\mathbf{Q}$ determines that of its negations and dual:

PROPOSITION 35.

(a) Outer negation reverses the direction of both RIGHT and LEFT MON.

(b) Inner negation reverses RIGHT MON but preserves LEFT MON.

(c) Dual-formation preserves RIGHT MON but reverses LEFT MON.

For example, from the monotonicity behaviour of one column of Table 1, we can infer that of all the other columns (for the binary quantifiers).

For doubly monotone quantifiers, we have the following pleasing result from van Benthem [1983c]. The proof is a nice demonstration of the strength and flexibility of the quantifiers constraints we are using.

THEOREM 36 (van Benthem). *Under CONSERV and VAR, the only doubly monotone quantifiers are those in the square of opposition.*

Proof. Suppose $\mathbf{Q}$ is $\downarrow$ *MON* $\downarrow$. We prove that $\mathbf{Q} = \mathbf{no}$; the theorem then follows from Proposition 35. Take a universe M and $A, B \subseteq M$. First assume that $A \cap B = \varnothing$. We claim that there is C such that $\mathbf{Q}_M AC$. This is immediate from *VAR* if $A \neq \varnothing$; otherwise, note that $\mathbf{Q}_M \varnothing\varnothing$ holds by $\downarrow$ *MON* $\downarrow$ and the fact that $\mathbf{Q}$ is non-trivial on M. By *MON* $\downarrow$ it then follows that $\mathbf{Q}_M A\varnothing$, i.e. $\mathbf{Q}_M AA \cap B$. Thus, by *CONSERV*, $\mathbf{Q}_M AB$. Conversely, suppose that $\mathbf{Q}_M AB$ holds. By $\downarrow$ *MON* $\downarrow$, $\mathbf{Q}_M A \cap BA \cap B$. But then $\mathbf{Q}_M A \cap BC$ holds for *all* $C \subseteq M$, since, for any such C, it suffices (by *CONSERV*) to show $\mathbf{Q}_M A \cap BA \cap B \cap C$, and this holds by *MON* $\downarrow$. Hence, *VAR* tells us that $A \cap B = \varnothing$, and the proof is finished. ■

For logical quantifiers, we can replace double monotonicity by *LEFT MON*:

THEOREM 37 (van Benthem). *The only logical and LEFT MON quantifiers satisfying VAR are the ones in the square of the opposition.*

A convenient method to prove this for *finite* universes (the case van Benthem considers) will be given in 4.2; actually, the result holds for all universes. Note the use of *VAR* here; without it, room is left for many other *LEFT MON* quantifiers, as is clear from the examples above.

Barwise and Cooper propose several universals involving monotonicity. One of them is the following:

($U10$) *Simple binary natural language quantifiers are either RIGHT MON or conjunctions of RIGHT MON quantifiers.*

Note that **exactly n** (which probably is simple) is the conjunction of the *RIGHT MON* **at least n** and **at most n**. This and other examples of neither *LEFT* nor *RIGHT MON* quantifiers suggest a weaker notion of monotonicity, which well be called *continuity*:

DEFINITION 38. A binary quantifier $\mathbf{Q}$ is

RIGHTCONT, if $\mathbf{Q}_M AB \& \mathbf{Q}_M AB' \&$
$\& B \subseteq B' \subseteq B'' \Rightarrow \mathbf{Q}_M AB'$,
LEFTCONT, if $\mathbf{Q}_M AB \& \mathbf{Q}_M A''B \&$
$\& A \subseteq A' \subseteq A'' \Rightarrow \mathbf{Q}_M A; B$.

Let us further call a quantifier *STRONG RIGHT* (*LEFT*) *CONT* if both it and its outer negation are *RIGHT* (*LEFT*) *CONT*. We have

RIGHT(*LEFT*)*MON* $\Rightarrow$
$\rightarrow$ *STRONG RIGHT (LEFT) CONT* $\Rightarrow$
$\Rightarrow$ *RIGHT*(*LEFT*)*CONT*.

None of the implications can be reversed: for example, **exactly n** is *RIGHT* (and *LEFT*) *CONT*, but not *STRONG RIGHT* (or *LEFT*) *CONT*.

Thijsse [1983] observes that the property of quantifiers identified in ($U10$) is in fact *RIGHT CONT*:

PROPOSITION 39. *A binary quantifiers is* RIGHT CONT *iff it is the conjunction of a* MON $\uparrow$ *and a* MON $\downarrow$ *quantifier.*

The proof is similar to the proof of Proposition 41(b) below.

Our use of the conservativity constraint on binary quantifiers gives the right and the left arguments quite different roles, so it is not surprising that right monotonicity and left monotonicity are very different properties. This is clear from Theorem 37, and will become even more apparent in Section 4.3. A further illustration of the difference is afforded by the following model-theoretic characterisation of the left monotonicity properties. Note first that any quantifier $\mathbf{Q}$ can be identified with a *class of structures*: in the binary case,

$$\mathbf{Q} = \{\langle M, A, B\rangle : \mathbf{Q}_M AB\}.$$

Call such a class *sub-closed* (*ext-closed*) if it is closed under substructures (extensions), and *inter-closed* if, whenever two structures, one a substructure of the other, are in $\mathbf{Q}$, then so is every structure 'between' these two. It is straightforward to verify that

PROPOSITION 40. *Under CONSERV and EXT, a binary quantifier is sub-closed (ext-closed, inter-closed) iff it is* $\downarrow$ MON *(*$\uparrow$ MON, LEFT CONT*).*

For *first-order definable* quantifiers, the semantic property of being subclosed has a well known syntactic counterpart, namely, definability by a *universal* sentence (cf. [Chang and Keisler, 1973, p. 128]). Thus, among first-order definable quantifiers satisfying *CONSERV* and *EXT*, the $\downarrow$ *MON* ones are precisely those definable by universal sentences. Corresponding results for $\uparrow$ *MON* and *LEFTCONT* quantifiers follow from the previous proposition and

PROPOSITION 41. *For any binary quantifier* $\mathbf{Q}$*,*

(a) $\mathbf{Q}$ *is ext-closed* $\Leftrightarrow$ $\neg\mathbf{Q}$ *is sub-closed,*

(b) $\mathbf{Q}$ *is inter-closed* $\Leftrightarrow$ $\mathbf{Q} = \mathbf{Q}' \wedge \mathbf{Q}''$*, for some sub-closed* $\mathbf{Q}'$ *and some ext-closed* $\mathbf{Q}''$*.*

Proof. (a) is obvious. As for (b), a conjunction of the sort indicated is clearly inter-closed. Conversely if $\mathbf{Q}$ is inter-closed, define

$$\mathbf{Q}'_M AB \Leftrightarrow \mathbf{Q}_{M'} A'B', \text{ for some extension } \langle M', A', B'\rangle \text{ of } \langle M, A, B\rangle,$$
$$\mathbf{Q}''_M AB \Leftrightarrow \mathbf{Q}_{M'} A'B', \text{ for some substructure } \langle M', A', B'\rangle \text{ of } \langle M, A, B\rangle;$$

then $\mathbf{Q}'$ and $\mathbf{Q}''$ are as desired. ■

Another syntactic characterisation of monotonicity from first-order logic is the following. Call a sentence $\phi(P)$, containing the unary P among its non-logical symbols, *upward monotone* (*in* P), if

$$\phi(P) \wedge \forall x(Px \rightarrow P'x) \rightarrow \phi(P')$$

is valid, and similarly for downward monotonicity. For example, sentences defining *LEFT* or *RIGHT MON* quantifiers will be monotone in certain predicate symbols. An occurrence of P in ϕ is said to be *positive* (*negative*), if it is within the scope of an even (odd) number of negations, when $\rightarrow$ and $\leftrightarrow$ have been eliminated. The next result is well known from first-order model theory (the proof is an application of Lyndon's interpolation theorem; cf. [Chang and Keisler, 1973, p. 90]).

PROPOSITION 42. *A first-order sentence $\phi(P)$ (which may contain other predicate symbols but no function or constant symbols) is upward (downward) monotone iff it is equivalent to a sentence where P occurs only positively (negatively).*

Monotonicity properties have been quite useful in describing and explaining linguistic phenomena; cf. [Barwise and Cooper, 1981; Keenan and Stavi, 1986], and, in connection with so-called polarity items, [Ladusaw, 1979; Zwarts, 1986]. We will have several further uses of monotonicity in Section 4. In mathematical logic, monotone quantifiers have been studied in model theory and recursion theory. The beginnings of the model theory for montone quantifiers will be given in Appendix B; further information can be found in [Barwise and Feferman, 1985]. On the more recursion-theoretic side, cf. for example, [Aczel, 1975] and [Barwise, 1978], and the references therein.

3.7 Partial and Definite Quantifiers

In 2.4.6 we mentioned that the number conditions belonging to the definites have been taken to indicate that the corresponding quantifiers are *partial*. This is the approach of Barwise and Cooper, who furthermore identify a semantic property of partial quantifiers, called *definiteness*, characteristic of the interpretation of the definites.[29]

Consider (in this subsection) binary quantifiers which are *partial in the first argument* (i.e. for certain A, $\mathbf{Q}_M AB$ may be *undefined* for all B). For example, the partial quantifier **the** coincides with the total **the** when $|A| = 1$, but is undefined when $|A| \neq 1$.

DEFINITION 43. $\mathbf{Q}$ is *definite*, if, for all M and all $A \subseteq M$ for which $\mathbf{Q}$ is defined, there is a non-empty set B_A such that, for all $B \subseteq M$, $\mathbf{Q}_M AB \Leftrightarrow B_A \subseteq B$.

The simple definites of 2.4.6 all have this property, when treated as partial quantifiers: e.g. for **the**, $B_A = A$ (or $B_A = X \cap A$ for some context set X), and for

[29] They consider (the singular) *the, both*, and *DET*s of the form *the n*, but not possessives.

John's, $B_A = P_{\mathrm{John}} \cap A$. That the use of partial quantifiers is necessary here follows from

PROPOSITION 44. *Under CONSERV, no definite quantifier is total.*

Proof. This follows from the fact that a definite and conservative quantifier must be undefined for $A = \varnothing$: suppose $\mathbf{Q}$ is defined for $\varnothing$ and consider $B_\varnothing$ that exists by definiteness. Since $B_\varnothing \subseteq B_\varnothing$ we have $\mathbf{Q}_M \varnothing B_\varnothing$ and thus, by *CONSERV*, $\mathbf{A}_M \varnothing\varnothing$. But then $B_\varnothing \subseteq \varnothing$, by definiteness, contradicting the stipulation that $B_\varnothing$ is non-empty. ■

In view of this proof it is natural to weaken the requirements in Definition 43 slightly. Call $\mathbf{Q}$ *universal*, if it is as in 43, *except* that $B_\varnothing$ is allowed to be empty (i.e. that B_A is required to be non-empty only when A is). All definite quantifiers are universal, but not conversely, since **all** is universal. This is indeed the prime example of a universal quantifier, as the next result shows.

THEOREM 45. *Suppose* $\mathbf{Q}$ *is logical. Then* $\mathbf{Q}$ *is universal iff* $\mathbf{Q} =$ **all** *whenever defined.*

Proof. If $\mathbf{Q}$ coincides with **all** whenever defined it is clearly universal (with $B_A = A$). Conversely, suppose $\mathbf{Q}$ is universal and defined for A. We need to show that $B_A = A$. If $A = \varnothing$ we get $B_A = \varnothing$ just as in the proof above. Suppose, then, that $A \neq \varnothing$. then $B_A \neq \varnothing$ by universality. Also, $B_A \subseteq A$; this follows from *CONSERV*, since $\mathbf{Q}_M A B_A$, whence $\mathbf{Q}_M AA \cap B_A$, and thus $B_A \subseteq A \cap B_A$ by universality. Now assume that $B_A \neq A$. Take $a \in B_A$ and $a' \in A - B_A$. Let f be a function which permutes a and a' but leaves everything else in M as it is. By *ISOM*, $\mathbf{Q}_M f[A] f[B_A]$, i.e. $\mathbf{Q}_M A(B - \{a\}) \cup \{a'\}$. Thus, by universality, $B_A \subseteq (B_A - \{a\}) \cup \{a'\}$. But this contradicts $a \in B_A$. ■

Thus the logical universal quantifiers, and in particular the definite ones, are just partial versions of **all**. This is one reason to restrict attention to total quantifiers, as we have done in preceding sections and shall continue to do in what follows. Another reason is that partial quantifiers make the model theory more cumbersome, and that many results for total quantifiers can rather easily be extended to the partial case by inserting phrases of the form 'whenever ... is defined' in suitable places.

Note finally that even if partial quantifiers are admitted in principle,, the alternative treatment of definites suggested in 2.4.7 makes it possible to propose the universal.

($U11$) *Natural language quantifiers are total,*

while still preserving the intuition that statements involving definites lack truth value when the corresponding number conditions are not met.

3.8 Finite Universes

Many *DET*s more or less presuppose that the N and *VP* denotations under consideration are finite sets. Examples are *more than half, 30 percent of, many*, but also *DET*s like *most, more ... than, fewer ... than*, where the interpretations we gave actually work for infinite sets as well. It seems that in many natural language contexts we can make the blanket assumption

FIN *Only finite universes are considered.*

For *DET*s like *infinitely many* or *all but finitely many*, on the other hand, infinite models seem to be needed. So our strategy will be to keep track of those results that need *FIN* and those that don't. Interestingly, it turns out that *FIN* is a very natural constraint for the quantifier theory in the next section, in the sense that it *simplifies results and proofs*. Most of the results have generalisations to the case when *FIN* is dropped, but the added information does not appear to be very exciting from a natural language point of view.

This should be contrasted with the situation in mathematical logic. There infinite sets are crucial, and finite models are often just a nuisance. Consider the effect *FIN* would have in classical model theory. Most standard methods of constructing models (compactness, ultraproducts, etc.) would become ineffective, and many of the usual logical questions would become pointless. For example, the four properties of logics mentioned in Section 1.6 lose their interest. This is clear for the Tarski and the Löwenheim property, and for compactness and completeness it follows from

PROPOSITION 46. *Under FIN, no logic is compact or complete.*

Proof. Under *FIN*, the set $\{\exists_{\geq n} x(x = x) : n = 1, 2, \ldots\}$ has no model, does *EL* (and hence all its extensions) fail to be compact. The statement about completeness follows from a result by Trakhtenbrot, by which the set of all finitely valid *EL*-sentences (i.e. the set of valid sentences under *FIN*) is not recursively enumerable. For any logic $L = L(\mathbf{Q}^i)_{i \in I}$, this set is the intersection of the set of finitely valid L-sentences with the (recursive) set of *EL*-sentences. It follows that the set of finitely valid L-sentences is not recursively enumerable. ■

Some standard logical questions remain, though. For example, we may still compare the *power of expression* of various logics under *FIN*, though some of the facts may change: we showed in 1.7 that $L(\mathbf{most}) < L(\mathbf{more})$ in general, but that $L(\mathbf{most}) \equiv L(\mathbf{more})$ under *FIN*. Likewise, *definability* issues are affected by *FIN*; for example, **all but finitely many** is not first-order definable in general, although it is trivially first-order definable under *FIN*.

It should be noted, however, that the main definability results in Section 1.7 (Theorem 10 and Corollary 11) continue to hold in the presence of *FIN*.

4 THEORY OF BINARY QUANTIFIERS

Binary quantifiers are the most common ones in natural language; they are also the most manageable relations, and we restrict attention to them from now on. A similar study of $(n + 1)$-ary quantifiers appears quite feasible, cf. [Keenan and Moss, 1985]. The important step is abandoning unary quantifiers: most of the results in this section have no counterpart for the unary case.

If nothing else is said, we assume in what follows that *all quantifiers involved are logical and satisfy NONTRIV*. Other constraints, such as *ACT, VAR* and *FIN*, will be stated explicitly.

As a consequence of the assumption that *EXT* holds, we can often skip reference to the universe M, and write

$$\mathbf{Q}AB$$

instead of $\mathbf{Q}_M AB$. More precisely, let $\mathbf{Q}AB$ mean that, for *some* M such that $A, B \subseteq M, \mathbf{Q}_M AB$. *EXT* then guarantees that this is well defined.

Most of the results in 4.1–5 below originate from [van Benthem, 1984a; van Benthem, 1983c].

4.1 Relational Behaviour

We have already encountered standard properties of binary relations, such as (ir)reflexivity (2.2.2) and symmetry (3.4), in the context of natural language quantification. A first start in quantifier theory is to exploit this perspective systematically. As we shall see, this turns out to be both rewarding in itself and useful for other purposes. Here are a few common properties of relations, and some quantifiers exemplifying them:

One project is to find informative characterisations of (logical) quantifiers having such properties. As for symmetry, two useful equivalent formulations were given in Proposition 30. To deal with the other properties, we first state a

LEMMA 47. *If* $\mathbf{Q}AB$ *holds, there exists* B' *such that* $\mathbf{Q}AB'$ *and* $\mathbf{Q}B'A$.

Proof. Choose B' such that $A \cap B = B' \cap A$ and $|A - B| = |B' - A|$ (this may require extending the original universe, which is permitted by *EXT*. Since $\mathbf{Q}AB$, we get $\mathbf{Q}AB'$ by *CONSERV*, and then $\mathbf{Q}B'A$ by *QUANT*. ■

Note the use of logicality here; the lemma fails if any of *CONSERV*, *EXT*, or *QUANT* are dropped. The following corollary is immediate (since we are assuming *NONTRIV*):

COROLLARY 48 (van Benthem). *There are no asymmetric quantifiers.*

A characterisation of antisymmetry is also forthcoming.

COROLLARY 49. $\mathbf{Q}$ *is antisymmetric iff* $\mathbf{Q}AB \Rightarrow A \subseteq B$.

Table 2.

Property	*Definition*	*Examples*
symmetry	$\mathbf{Q}AB \Rightarrow \mathbf{Q}BA$	**some, no, at least n, at most n, exactly n, between n and m**
antisymmetry	$\mathbf{Q}AB\&\mathbf{Q}BA \Rightarrow A = B$	**all**
asymmetry	$\mathbf{Q}AB \Rightarrow \neg\mathbf{Q}BA$	-
reflexivity	$\mathbf{Q}AA$	**all, at lest half, all but finitely many**
quasireflexivity	$\mathbf{Q}AB \twoheadrightarrow \mathbf{Q}AA$	**some, most, at least n**
weak reflexivity	$\mathbf{Q}AB \Rightarrow \mathbf{Q}BB$	**some, most, at least n**
quasiuniversality	$\mathbf{Q}AA \Rightarrow \mathbf{Q}AB$	**no, not all, all but n**
irreflexivity	$\neg\mathbf{Q}AA$	**not all, all but n**
linearity	$\mathbf{Q}AB \vee \mathbf{Q}BA \vee A = B$	**not all**
transitivity	$\mathbf{Q}AB\&\mathbf{Q}BC \Rightarrow \mathbf{Q}AC$	**all, but finitely many**
circularity	$\mathbf{Q}AB\&\mathbf{Q}BC \Rightarrow \mathbf{Q}CA$	-
euclidity	$\mathbf{Q}AB\&\mathbf{Q}AC \Rightarrow \mathbf{Q}BC$	-
antieuclidity	$\mathbf{Q}AB\&\mathbf{Q}CB \Rightarrow \mathbf{Q}AC$	?

Proof. If the condition holds, $\mathbf{Q}$ is clearly antisymmetric. Conversely, if $\mathbf{Q}$ is antisymmetric and $\mathbf{Q}AB$ holds, take B' as in the proof of Lemma 47. Thus $A = B'$ by antisymmetry, and $|A - B| = |B' - A| = 0$, i.e. $A \subseteq B$. ■

This also gives a characterisation of linearity, since $\mathbf{Q}$ is linear iff $\neg\mathbf{Q}$ is antisymmetric. As to the reflexivity properties and quasiuniversality, their main interest is in combination with other properties, as we shall see. The following consequences of Lemma 47 may nevertheless be noted:

COROLLARY 50. *Weak reflexivity implies quasireflexivity.*

This leaves only the properties in Table 2 involving *three* set variables. The '-' signs here are explained by the following results from van Benthem [1984a].

THEOREM 51 (van Benthem). *There are no Euclidean quantifiers.*

We omit the proof, but show how to obtain the following corollary with the aid of Lemma 47

COROLLARY 52 (van Benthem). *There are no circular quantifiers.*

Proof. Suppose $\mathbf{Q}$ is circular. If $\mathbf{Q}AB$, take B' as in Lemma 47. By circularity, $\mathbf{Q}AA$. Thus, $\mathbf{Q}AB \Rightarrow \mathbf{Q}AA\&\mathbf{Q}AB \Rightarrow \mathbf{Q}BA$ (again by circularity), i.e. $\mathbf{Q}$ is symmetric. But it is easy to see that a circular and symmetric quantifier is Euclidean, contradicting the theorem. ■

Actually, some of these results, e.g. Corollary 48 and Theorem 51, were first proposed as semantic universals, based on empirical evidence (Frans Zwarts). Only later was it realised that they are consequences of more fundamental properties of quantifiers. This provided a first illustration of the potential usefulness of quantifier theory for linguistic explanation.

We left a question mark for antieuclidity in Table 2. Here is an example though: $\mathbf{Q}AB \Leftrightarrow |A| = n$. The following result from [Westerståhl, 1984] explains the situation.

THEOREM 53. $\mathbf{Q}$ *is antiEuclidean iff* $\mathbf{Q}AB \Rightarrow \mathbf{Q}AC$ *(for all* A, B, C*).*

Two corollaries follow easily:

COROLLARY 54. $\mathbf{Q}$ *is antiEuclidean iff there is a class* X *of cardinal numbers such that* $\mathbf{Q}AB \Leftrightarrow |A| \in X$.

COROLLARY 55 (Zwarts). *Under* VAR *there are no antiEuclidean quantifiers.*

Thus antiEuclidean quantifiers put no condition at all on the *second* argument, i.e. the *VP* denotation. It seems safe to conclude that there are no antiEuclidean natural language quantifiers.

Finally, consider transitivity. Here are some examples of transitive quantifiers:

(a) **all, all but finitely many,**

(b) $\mathbf{all}_e AB \Leftrightarrow \varnothing \neq A \subseteq B$ (**all** with existential import; cf. 3.4)

(c) $\mathbf{all}_n AB \Leftrightarrow A \subseteq B \vee |A| < n$ ($n \geqq 1$; note that $\mathbf{all}_1 = \mathbf{all}$)

(d) any antiEuclidean quantifier (by Theorem 53)

(e) $\mathbf{Q}AB \Leftrightarrow (A \subseteq B \& |A| \geqq 5) \vee |A| = 3$.

Let us check (e): suppose $\mathbf{Q}AB$ and $\mathbf{Q}BC$. In case $|A| = 3$ we get $\mathbf{Q}AC$ automatically, so suppose $A \subseteq B \& |A| \geqq 5$. But then $|B| \neq 3$, so we must have $B \subseteq C \& |B| \geqq 5$, whence $A \subseteq C \& |A| \geqq 5$, i.e. $\mathbf{Q}AC$. ■

The following characterisation of transitivity from [Westerståhl, 1984] depends essentially on *FIN*. It shows that (e) above is in a sense the typical case. If X, Y are sets of natural numbers, let $X < Y$ mean that every number in X is smaller than every number in Y; this is taken to hold trivially if X or Y are empty.

THEOREM 56 (*FIN*). $\mathbf{Q}$ *is transitive iff there are sets* X, Y *of natural numbers such that* $X < Y$ *and* $\mathbf{Q}AB \Leftrightarrow |A| \in X \vee (A \subseteq B \& |A| \in Y)$.

The proof combines a result from [van Benthem, 1984a] with techniques that will be introduced in 4.2 below. Note that the transitive **all but finitely many** fails to satisfy the condition in the theorem, if infinite universes are allowed. The next corollary shows that *VAR* has drastic effects on transitivity.

COROLLARY 57 (*FIN*). *Under* VAR *the only transitive quantifiers are* **all** *and* $\mathbf{all}_e$.

Proof. This follows from the observation that *VAR* implies that either $X = \varnothing$ and $Y = N$, or $X = \{0\}$ and $Y = N = \{0\}$ in the theorem. ■

Having thus looked at single properties of quantifiers, we can go on to *combinations* of such properties. For example, using Theorem 51 and Proposition 30 we obtain the

COROLLARY 58. *No quantifiers are both*

(a) symmetric and transitive,

(b) symmetric and antiEuclidean,

(c) symmetric and (ir)reflexive,

(d) quasiuniversal and reflexive.

Reflexivity often has strong effects in combination with other properties. Note that, if $\mathbf{Q}$ is reflexive, $A \subseteq B \Rightarrow \mathbf{Q}AB$ (by *CONSERV*). From this and Corollary 49 we immediately get

COROLLARY 59. *The only reflexive and antisymmetric quantifier is* **all**.[30]

Furthermore, it is not hard to see that reflexivity together with the condition in Theorem 56 implies that, for some $n \geqq 1$, $X = \{0, \ldots, n-1\}$ and $Y = \{k : k \geqq n\}$. This gives

COROLLARY 60 (van Benthem (*FIN*)). *The only reflexive and transitive quantifies are* $\mathbf{all}_n$, *for* $n \geqq 1$.

Again, **all but finitely many** is a counterexample if *FIN* is dropped.

COROLLARY 61 (*FIN*). *Under ACT, the only reflexive and transitive quantifier is* **all**.

Proof. Suppose that $\mathbf{Q} = \mathbf{all}_n$, for some $n \geqq 2$. Let M be a universe with exactly one element. It follows that $\mathbf{Q}$ is trivial on M, contradicting *ACT*. ■

The next result connects our simple properties of relations with the monotonicity properties of Section 3.6.

THEOREM 62 (Zwarts).

(a) If $\mathbf{Q}$ *is reflexive and transitive, then* $\mathbf{Q}$ *is* $\downarrow$ MON $\uparrow$.

(b) If $\mathbf{Q}$ *is symmetric, then*

(i) $\mathbf{Q}$ *is quasireflexive iff* $\mathbf{Q}$ *is* MON $\uparrow$,

(ii) $\mathbf{Q}$ *is quasiuniversal iff* $\mathbf{Q}$ *is* MON $\downarrow$.

[30] Actually, only *CONSERV* is needed for this result [van Benthem, 1984a].

Proof. We prove (a); (b) is similar. If $\mathbf{Q}AB$ and $A' \subseteq A$, then $\mathbf{Q}A'; A$, by reflexivity and *CONSERV*, and hence $\mathbf{Q}A; B$ by transitivity. Similarly, if $\mathbf{Q}AB$ and $B \subseteq B'$, then $\mathbf{Q}BB'$ and hence $\mathbf{Q}AB'$. ■

From this and Theorem 36 we get the following variant of Corollary 61.

COROLLARY 63. *Suppose that* $\mathbf{Q}$ *satisfies* CONSERV *and* VAR *(but not necessarily* EXT *or QUANT), and is reflexive and transitive. Then* $\mathbf{Q} = \mathbf{all}$.

Proof. It suffices to note that neither Theorem 36 nor Theorem 62 uses *EXT* or *QUANT*. ■

Instead of characterising properties in terms of which quantifiers satisfy them, one may turn the question around and ask for characterisations of our most common quantifiers in terms of their properties. For the quantifier **all** and its variants, such characterisations were in fact obtained in Corollaries 57, 59–61, and 63. We end by giving a corresponding result for **some**. Let, for each cardinal κ, $\mathbf{some}_\kappa$ be the quantifier **at least** κ, i.e.

$$\mathbf{some}_\kappa AB \Leftrightarrow |A \cap B| \geqq \kappa$$

(so $\mathbf{some}_1 = \mathbf{some}$).

THEOREM 64 (van Benthem). $\mathbf{Q}$ *is symmetric and quasireflexive iff* $\mathbf{Q} = \mathbf{some}_\kappa$, *for some* $\kappa \geqq 1$.

A proof will be given in Section 4.2. The following corollary is obtained similarly to Corollary 61.

COROLLARY 65. *Under ACT, the only symmetric and quasireflexive quantifier is* **some**.

4.2 Quantifiers in the Number Tree

By Proposition 24, each binary logical quantifier $\mathbf{Q}$ can be identified with a binary relation between cardinal numbers. We use the same notation for this relation, which is thus defined by

(1) $\mathbf{Q}xy \Leftrightarrow$ for some A, B with $|A - B| = x$ and $|A \cap B| = y$, $\mathbf{Q}AB$.

Inversely, given any binary relation $\mathbf{Q}$ between cardinal numbers, we get the corresponding logical quantifier by

(2) $\mathbf{Q}AB \Leftrightarrow \mathbf{Q}|A - B|\ |A \cap B|$.

With (1) and (2) we can switch back and forth between a *set-theoretic* and a *number-theoretic perspective* on quantifiers. The latter perspective is the subject of the present subsection.

Here are the number-theoretic versions of a few well known quantifiers:

all $xy \Leftrightarrow x = 0$,
no $xy \Leftrightarrow y = 0$,
some$_n xy \Leftrightarrow y \geqq n$,
all$_n xy \Leftrightarrow y = 0 \vee x + y < n$,
most $xy \Leftrightarrow y < x$,
infinitely many $xy \Leftrightarrow y$ is infinite,
all but finitely many $xy \Leftrightarrow x$ is finite.

Properties of quantifiers also have their number-theoretic versions. In the case of universal properties, such as those in Table 2, there is a simple translation from the set-theoretic to the number-theoretic framework. Details can be found in [Westerståhl, 1984]; here we just consider a few examples. If two sets A, B are involved, let x correspond to $|A - B|$, y to $|A \cap B|$, and z to $|B - A|$. Then, for example,

(3) *quasireflexivity* is the property: $\mathbf{Q}xy \Rightarrow \mathbf{Q}0x + y$ (for all x, y),

(4) *symmetry* is the property: $\mathbf{Q}xy \Rightarrow \mathbf{Q}zy$ (for all x, y, z), or, equivalently, $\mathbf{Q}xy \Leftrightarrow \mathbf{Q}0y$ (for all x, y);

the last equivalence follows from Proposition 30 (it is also easy to see directly).

Sometimes proofs are simpler to carry out in the number-theoretic framework. This holds for several of the results in 4.1, in particular Theorems 53 and 56. As an illustration, we give the following

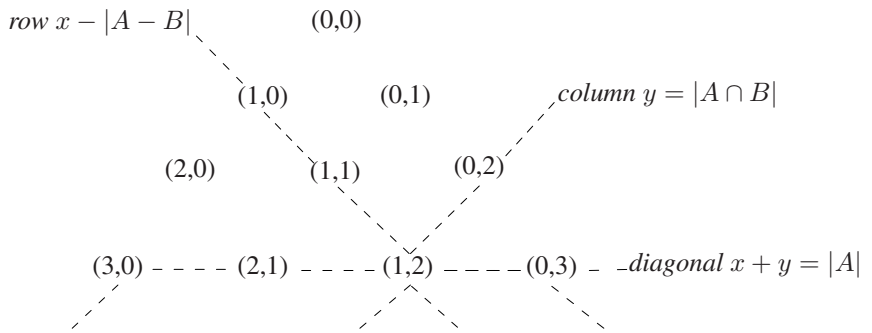
Proof.[of Theorem 64] Let κ be the least cardinal x such that $\mathbf{Q}0x, \kappa$ exists, by *NONTRIV* and (4). Also, $\kappa > 0$; otherwise, for any x, y, we get $\mathbf{Q}y0$ (from $\mathbf{Q}00$ by (4)), whence $\mathbf{Q}0y$ (by (3)), and so $\mathbf{Q}xy$ (by (4)), contradicting *NONTRIV*. We claim that $\mathbf{Q} = \mathbf{some}_\kappa$. Clearly, $\mathbf{Q}xy$ implies $y \geqq \kappa$, by 94). Conversely, given x, y such that $y \geqq \kappa$, take x' such that $\kappa + x' = y$. By (4) and the definition of κ, $\mathbf{Q}x'\kappa$. Thus, by (3), $\mathbf{Q}0x' + \kappa$ i.e. $\mathbf{Q}0y$ so $\mathbf{Q}xy$ by (4). ■

An operation that becomes nicely represented in the number-theoretic framework is *inner negation*, since we have

PROPOSITION 66. $(\mathbf{Q}\neg)xy \Leftrightarrow \mathbf{Q}yx$.

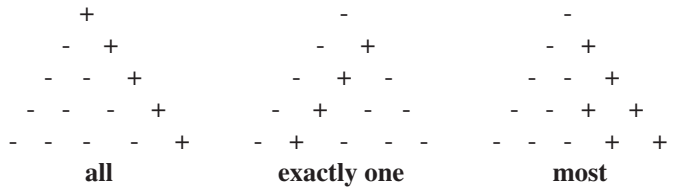
The number-theoretic perspective becomes particularly attractive if *FIN* is assumed. Quantifiers are then subsets of N^2. N^2 can be represented as a *number tree*, where each point (x, y) has two *immediate successors* $(x+1, y)$ and $x, y+1)$, which in turn are the *immediate predecessors* of the point $(x + 1, y + 1)$:[31]

[31] Without *FIN* one may represent logical quantifiers as subsets of $Card^2$ (*Card* = the class of cardinal numbers). This is not as easy to visualise as N^2. For example, diagonals and columns get mixed up:$(0, \aleph_0), (1, \aleph_0), \ldots$ are in the column given by $\aleph_0$, but also in the diagonal $\{(x, y) : x + y = \aleph\}$.



Quantifiers and their properties can be visualised in the number tree, and proofs can often be carried out directly in it. For an illustrative example, the reader is invited to carry out the above proof of theorem 64 in the number tree (assuming *FIN*). Note that symmetry (quasi-reflexivity) means that if a point is in **Q** then so are all the points on the column (so is the rightmost point on the diagonal) though it. Another illustration, also left to the reader, is the proof of Proposition 33 in the number tree.

When representing a quantifier **Q** in the tree it is often practical to write a '+' on the points in **Q** and a '-' on the other points. For example,



With this technique we can give our *non-triviality* conditions the following perspicuous formulations (we assume *FIN* for the rest of this subsection):

(5) *NONTRIV* $\Leftrightarrow$ there is at least one + and one -in the tree,

(6) *ACT* $\Leftrightarrow$ there is at least one + and one - *in the top triangle* (0,0),(1,0), (0,1),

(7) *VAR* $\Leftrightarrow$ there is at least one +and one - *on each diagonal* (except (0,0)).

This illustrates that *VAR* is a much stronger assumption than *ACT*, i.e. that the universal $(U9)$ in 3.5 really has content.

Monotonicity properties turn out to be particularly suited to number tree representation. Beginning with the *RIGHT* monotonicity properties, we can easily verify that

(8) *MON* $\uparrow \Leftrightarrow$ each + fills the diagonal to its right with +s,

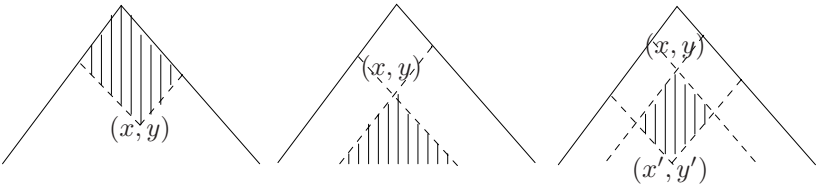
(9) *MON* $\downarrow \Leftrightarrow$ each + fills the diagonal to its left with +s,

(10) *RIGHT CONT* $\Leftrightarrow$ between two +s on a diagonal there are only +s.

Also observe that *STRONG RIGHT CONT*, i.e. *RIGHT CONT* for both $\mathbf{Q}$ and $\neg\mathbf{Q}$, amounts to (10) together with the same condition with '+' replaced by '-'. It follows that

(11) *STRONG RIGHT CONT* $\Leftrightarrow$ on each diagonal there is at most one change of sign.

The *LEFT* monotonicity properties can be illustrated as follows:

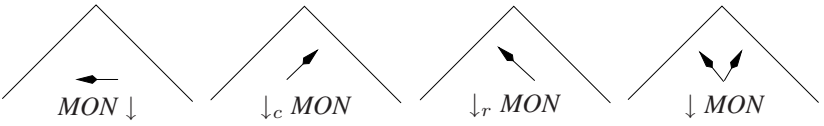


I.e. if (x, y) (and (x', y')) is in $\mathbf{Q}$ then so are all the points in the shaded area.

Working in the number tree, we can introduce several variants of the above monotonicity properties. Define $\uparrow_c\ MON, \downarrow_c\ MON, LEFT_cCONT$, and *STRONG LEFT*$_c$ *CONT* by replacing in (8)–(11), respectively 'diagonal' with 'column', and do the same for $\uparrow_r\ MON, \downarrow_r\ MON, LEFT_rCONT$, and *STRONG LEFT*$_r$ *CONT*, replacing 'diagonal' with 'row'. The terminology is motivated by the fact that

(12) $\uparrow_c MON \Leftrightarrow (\mathbf{Q}AB\&A' \subseteq A\&A \cap B = A' \cap B \Rightarrow \mathbf{Q}A'B)$,

and similarly for the other properties; in other words, they are as the previous *LEFT* properties, only we keep $A \cap B$ fixed in the 'c' case, and $A - B$ fixed in the 'r' case. To make the intuitive picture clear, her is yet another way to illustrate the downward monotone properties we have so far encountered:



In the tree it is easy to check whether particular quantifiers have such properties. For example, it is clear from the above illustrations that **most** is *MON* $\uparrow$ and $\downarrow_c$ *MON*, but not $\downarrow_r$ *MON*. It is also clear that

(13) $\uparrow$ *MON* $\Leftrightarrow\uparrow_c$ *MON*& $\uparrow_r$ *MON*,

(14) $\downarrow$ *MON* $\Leftrightarrow\downarrow_c$ *MON*& $\downarrow_r$ *MON*.

The corresponding statement for *CONT* fails, however (as can also be seen from the tree).

An interesting application of 'tree techniques' is given in [van Benthem, 1983c] to an idea in [Barwise and Cooper, 1981] concerning how hard it is (psychologically) to 'process' (verify or falsify) quantified statements. Barwise and Cooper speculated that quantifiers with monotonicity properties were easier to process and would therefore be preferred in natural language. Now, verifying a sentence of the form *all* AB in a universe with n elements takes n observations, and falsifying it takes at least 1 observation. If we look at *most* AB instead (and suppose that n is even for simplicity), the least possible number of observations it takes to verify it is $n/2 + 1$, and the corresponding number for falsification is $n/2$. In both cases the sum is $n + 1$. This holds for many basic quantifiers, but not all: e.g. *exactly one* AB requires n observations for verification and 2 for falsification.

van Benthem defines, with reference to the number tree, $\mathbf{Q}$ to be of *minimal count complexity* if, on each universe with n elements (this corresponds to the finite top triangle of the tree with the diagonal $x + y = n$ as base), there is a minimal confirmation pair (x_1, y_1) and a minimal refutation pair $(x_2, y_2)(x_i + y_i \leqq n)$ such that every pair (x, y) on the diagonal $x + y = n$ is determined by them:

$$x \geqq x_1 \& y \geqq y_1 \Rightarrow \mathbf{Q}xy,$$
$$x \geqq x_2 \& y \geqq y_2 \Rightarrow \neg\mathbf{Q}xy$$

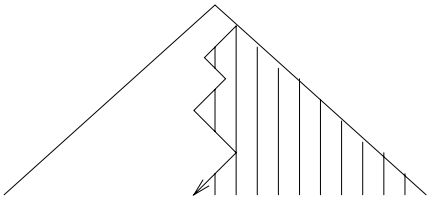
One can verify that $x_1 + y_1 + x_2 + y_2 = n + 1$, and thus that **all** and **most** are of minimal count complexity, but not **exactly one**.

Now consider the very strong continuity property:

$$\begin{aligned} SUPER\ CONT =_{\mathrm{df}}\ & STRONG\ RIGHT\ CONT\& \\ & \&STRONG\ LEFT_c CONT\& \\ & \&STRONG\ LEFT_r CONT. \end{aligned}$$

In other words, *SUPER CONT* means that there are no changes of sign in any of the three main directions in the number tree. It can be seen that the *SUPER CONT* quantifiers are precisely those determined by a *branch* in the tree (which can start anywhere on the edges; not necessarily at the top) with the property that, going downward, it always contains one of the immediate successors of each point on it:

The connection with count complexity is now the following:



THEOREM 67 (van Benthem). *(FIN) Under ACT,* $\mathbf{Q}$ *is of minimal count complexity iff it is SUPER CONT.*

The proof of this consists simply in showing that the two combinatorial descriptions give the same tree pattern.

From the above description of *SUPER CONT* one also obtains the following results:

PROPOSITION 68. *SUPER CONT* $\Rightarrow$ *RIGHT MON*.[32]

PROPOSITION 69. *There are uncountably many SUPER CONT logical quantifiers (even under FIN).*

What is the relation between *SUPER CONT* and *LEFT CONT*? Using the tree it is easy to see that neither property implies the other. In the next subsection we shall find, moreover, that there are only countably many *LEFT CONT* logical quantifiers, under *FIN*.

4.3 First-order Definability and Monotonicity

We shall prove a theorem characterising the first-order definable quantifiers in terms of monotonicity, under *FIN*. The most general form of the result has nothing directly to do with logicality, so we begin by assuming that $\mathbf{Q}$ is an arbitrary k-ary quantifier ($K \geqq 1$). We noted in 3.6 that $\mathbf{Q}$ can be identified with the class of structures $\langle M, A_0, \ldots, A_{k-1}\rangle$ such that $\mathbf{Q}_M A_0, \ldots, A_{k-1}$, and we defined the properties of being sub-closed, ext-closed, and inter-closed for classes of structures.

The key to the result is the following lemma from [van Benthem, 1984a]:

LEMMA 70 (van Benthem). *(FIN) Suppose* $\mathbf{K}$ *is a class of (finite) structures which is definable in EL by a set of monadic universal sentences. Then* $\mathbf{K}$ *is definable already by one such sentence.*

THEOREM 71. *(FIN)* $\mathbf{Q}$ *is first-order definable iff there are interclosed quantifiers* $\mathbf{Q}_1, \ldots, \mathbf{Q}_m$ *satisfying ISOM such that* $\mathbf{Q} = \mathbf{Q}_1 \vee \ldots \vee \mathbf{Q}_m$.

Proof. Suppose first $\mathbf{Q}$ is a disjunction of this kind. By Proposition 41, each $\mathbf{Q}_i$ can be written $\neg\mathbf{Q}'_i \wedge \mathbf{Q}''_i$, where $\mathbf{Q}'_i$ and $\mathbf{Q}''_i$ are sub-closed. Moreover, it easily follows from the proof of that proposition that both $\mathbf{Q}'_i$ and $\mathbf{Q}''_i$ satisfy *ISOM* if $\mathbf{Q}_i$ does. Thus it will suffice to show that every sub-closed quantifier satisfying *ISOM* is first-order definable. Assume, then, that $\mathbf{Q}$ has these properties. Under *FIN*, *any* class of structures closed under isomorphism is definable by a *set* of *EL*-sentences, by a standard argument: a finite structure can be completely described (up to isomorphism) by *one EL*-sentence, and the relevant set consists of all negated descriptions of models *not* in the class. If the class is in addition sub-closed, a variant of this argument shows that the sentences can be taken universal (one takes the negations of the existentially quantified diagrams of structures not in the class).[33] Since in our case the class is also monadic, $\mathbf{Q}$ is first-order definable by Lemma 70.

[32] This does not need *FIN*.

[33] This observation is also from [van Benthem, 1984a]. For the notion of a diagram, cf. [**?**, p. 68].

Now suppose $\mathbf{Q}$ is definable by an *EL*-sentence $\psi = \psi(P_0, \ldots, P_{k-1})$. By Corollary 11 (with $L = EL$), there is a natural number n such that $\mathbf{Q}$ is closed under the relation $\approx_n$ (cf. Section 1.7). Consider sentences expressing conditions

$$|P_s^{\mathbf{M}}| = i,$$

for some $i < n$, *or*

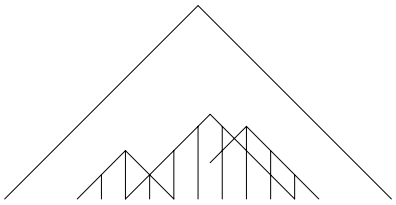
$$|P_s^{\mathbf{M}}| \geqq n.$$

It follows that any conjunction of such sentences where s runs through all the functions from k to 2, is a complete description of a model $\langle M, A_0, \ldots, A_{k-1}\rangle$, as far as $\mathbf{Q}$ is concerned. There are finitely many such descriptions, and ψ must be equivalent to the disjunction of all complete descriptions of structures in $\mathbf{Q}$. Moreover, each disjunct defines a quantifier, which, by the form of the definition, is easily seen to be inter-closed. Since any *EL*-definable quantifier satisfies *ISOM*, the theorem is proved. ■

Returning now to the case of binary logical quantifiers, we get from the theorem and Proposition 40 that

COROLLARY 72. *(FIN) If* $\mathbf{Q}$ *is binary and logical, then* $\mathbf{Q}$ *is first-order definable iff* $\mathbf{Q}$ *is a finite disjunction of LEFT CONT (binary and logical) quantifiers.*

There is a simpler direct proof of the corollary. This is because we can work in the number tree. In one direction, it suffices to show that $\uparrow$ *MON* quantifiers are first-order definable. If $\mathbf{Q}$ is $\uparrow$ *MON*, each point in $\mathbf{Q}$ generates an infinite downward triangle. From a given triangle within $\mathbf{Q}$, only finitely many steps can be taken towards the edges of the tree. It follows that $\mathbf{Q}$ is a finite union of such triangles,



and therefore clearly first-order definable. The proof in the converse direction, using Corollary 11, also becomes simpler in the number tree.

Corollary 72 shows, once more, that the *LEFT* monotonicity properties are much stronger than the *RIGHT* ones, due to the special role *CONSERV* gives to the left argument of a quantifier. In particular, there are only denumerably many *LEFT CONT* logical quantifiers (under *FIN*); this should be contrasted with Proposition 69.

Note that *FIN* is essential here. For example, **at most finitely many** is $\downarrow$ *MON* but not definable by any first-order sentence (or set of such sentences).[34]

Definability results such as these have not only logical interest: they also tell us something about the extent to which a certain logic — first order logic in this case — is adequate for natural language semantics. Of course, we knew already that first-order logic is not adequate, e.g. by the non-definability of **most**, but Corollary 72 places such isolated facts in a wider perspective.

The results here concern definability in the set-theoretic framework for quantifiers. What about *number-theoretic definability* (for logical quantifiers, under *FIN*)? Here we should consider formulas $\phi(x, y)$ in some suitable arithmetical language, containing at least the individual constant 0 and the unary successor function symbol S (and hence the *numerals* $\mathbf{0} = 0, \mathbf{1} = S0, \mathbf{2} = SS0$, etc.). Then ϕ *defines* $\mathbf{Q}$ iff, for all m, n,

$$\mathbf{Q}mn \Leftrightarrow \langle N, 0, S, \ldots\rangle \vDash \phi(\mathbf{m}, \mathbf{n}).$$

Examples of definable quantifiers, some of which in languages with the relation $<$ or the operation $+$, were given at the beginning of Section 4.2. Now which arithmetical definability notion corresponds to first-order definability in the set-theoretic sense? Notice first that even the simple formula

$$x = y$$

defines a non-first-order definable quantifiers, namely, **exactly half**. However, let the *pure number formulas* be those formulas in the language $\{0, S\}$ obtained from atomic formulas of the form

$$x = \mathbf{n}$$

by closing under Boolean connectives. Clearly every pure number formula with variables among x, y defines a first-order definable quantifier. But also conversely, for it can be seen by inspecting more closely the proofs of Theorem 71 and Corollary 72 that every first-order definable quantifier is in fact a Boolean combination of quantifiers of the form **at most n** and **all but at most n**, and the former, for example, is defined by the pure number formula

$$y = \mathbf{0} \vee \ldots \vee y = \mathbf{n}.$$

Thus, we have the

COROLLARY 73. *(FIN)* $\mathbf{Q}$ *is first-order definable iff* $\mathbf{Q}$ *is arithmetically defined by some pure number formula.*

This of course raises new definability questions. Which quantifies are defined by arbitrary formulas in $\{0, S\}$? Which are defined by formulas in $\{0, S, +\}$? It

[34] Michał Krynicki has observed (private communication) that, without *FIN*, *LEFTCONT* quantifiers are definable in logic with the cardinality quantifiers $\mathbf{Q}_\alpha$ (Section 1.3).

can be seen that **most** belongs to the second, but not the first, class. These questions are studied in connection with *computational complexity* in van Benthem [1985; 1987a]. He shows, among other things, that the second class of quantifiers mentioned above consists precisely of those computable by a push-down automaton (under *FIN*). He also characterises the first-order definable quantifiers computationally, namely, as those computable by a certain type of finite-state machine. This illustrates another aspect of the interest of definability questions: *classification* of quantifiers w.r.t. various notions of complexity. For the relevant definitions, and for several other interesting results along the same lines, we must refer to the two papers by van Benthem mentioned above.

4.4 Logical Constants

Clearly not all the $2^{2^{\aleph_0}}$ logical quantifiers deserve the title *logical constant*. We have already presented conditions that severely restrict the range of quantifiers. For example, *LEFT MON* plus *VAR* leaves only the quantifiers in the square of opposition (3.6). But there is no immediate reason why these two constraints should apply to logical constants. In this subsection, we look at some conditions which can be taken to have an independent connection with logical constanthood.

One idea seems natural enough, namely, that quantifiers that are logical constants should be *simple* natural language quantifiers (Section 2.4). Thus, the semantic universals holding for simple quantifiers apply to them. It follows that they should be logical (i.e. obey *CONSERV, EXT*, and *QUANT*) and satisfy *NONTRIV* and *RIGHT CONST* (by (U10) and Proposition 39).

As for constraints specifically related to logical constanthood, we will concentrate on one rather strong property often claimed to be characteristic of logical constants, namely, that *they do not distinguish cardinal numbers*. The idea is that such distinctions belong to mathematics, not logic. We will consider two rather different ways of making this idea precise.

FIN is used in what follows, so that we can argue in the number tree. It is possible, however, to generalise the results (with suitable changes) to infinite universes.

The first version of the above idea goes back to Mostowski [1957], although he only applied it to the infinite cardinalities. Given $\mathbf{Q}AB$, the relevant cardinality here is that of the universe, or in our case, by *CONSERV* and *EXT*, that of A. We must of course separate 0 from the other cardinalities, since distinguishing non-zero numbers from 0 is precisely what basic quantifiers such as **some** and **all** do. With these observations, we can transplant Mostowski's idea to the finite case as follows:

DEFINITION 74. Suppose $m, n > 0$. $\mathbf{Q}$ does not distinguish m and n if

(a) $\mathbf{Q}m0 \Leftrightarrow \mathbf{Q}n0$,

(b) $\mathbf{Q}0m \Leftrightarrow \mathbf{Q}0n$,

(c) if $x_1 + y_1 = m$ and $x_2 + y_2 = n$, where $x_i, y_i > 0$, then $\mathbf{Q}x_1y_1 \Leftrightarrow \mathbf{Q}x_2y_2$.

For example, **at least k** does not distinguish any $m, n < k$, but distinguishes all m, n for which at least one is $\geqq k$.

Note that no restriction at all is put on the point $(0, 0)$. To avoid trivial complications n the next result, we shall restrict attention to the number tree *minus* $(0, 0)$ (we write '-0' to indicate this). Also, we replace, in this subsection, *NONTRIV* by the slightly stronger condition that in the tree *minus*$(0, 0)$, there is at least one $+$ and one $-$.

It is not surprising that the present logicality constraint has rather drastic effects on the range of quantifiers:

THEOREM 75 (*FIN*, -0). *Suppose that* $\mathbf{Q}$ *does not distinguish any pair of non-zero natural numbers and satisfies* RIGHT CONT. *Then* $\mathbf{Q}$ *is one of the quantifiers* **some, no, all, not all**, *and* **some but not all**.

Proof. There are four possible patterns for the top triangle minus $(0, 0)$.
Case 1: $++$. By the cardinality property and *RIGHT CONT*, this puts a $+$ everywhere, contradicting (our present version of) *NONTRIV*. *Case 2:* $+-$. Then the left edge will have only $+$, and the right edge only $-$. For the remaining interior triangle, there are two possibilities, giving either **no** or **not all**. *Case 3:* $-+$. This is symmetric to Case 2 an gives **some** and **all**. *Case 4:* $--$. Besides the trivial case with only $-$, there is the case with $+$ in the interior triangle and $-$ on both edges, i.e. **some but not all**. ■

We may note that *VAR*, or *STRONG RIGHT CONT*, will exclude **some but not all**, but it is not clear that we have to assume any of these. (On the other hand, it could be argued that one interpretation of the *DET some*, especially when focused or stressed, *is* **some but not all**.)

The second version of the requirement that logical constants do not distinguish cardinal numbers is from [van Benthem, 1984a]. Here the idea is that no point in the tree is special: you always *proceed downward in the same way*. Proceeding one step downward can be regarded as a thought experiment, whereby one, given A and B, adds one element to $A - B$ or $A \cup B$. The condition is then that the outcome is *uniform* in the tree, i.e. that it does not depend on the number of elements in these sets (the point $(0, 0)$ need not be excluded here, although it can be):

UNIF *The sign of any point in the tree determines the sign of its two immediate successors.*

THEOREM 76 (van Benthem (*FIN*)). *The UNIF and* RIGHT CONT *quantifiers are precisely* **some, no, all, not all** *and the quantifiers* $|A|$ **is even** *and* $|A|$ **is odd**.

Proof. Again a simple tree argument suffices. There are eight top triangles to consider; let us look at two. First consider $\overset{-}{-}+$. Here the $-$ successors are determined, but for the right $+$ successor there is a choice, and we get two patterns

The first of these is excluded by *RIGHT CONT*, and the second is **some**.

Now consider the top triangle $\overset{+}{-}-$. Here the $-$successors are either both $-$ or both $+$. In the first case we get only $-$ in the rest of the tree, contradicting *NONTRIV*. In the second case we get $|A|$ **is even**. the other cases are similar. ■

The last two quantifiers in the theorem are not natural language quantifiers and should be excluded somehow. The following slight strengthening of *NONTRIV* would suffice:

NONTRIV* *On some diagonal in the tree, there is at least one* $+$ *and at least one* $-$.

In fact, it seems that we may safely replace *NONTRIV* by *NONTRIV** in the universal $(U8)$ in 3.5.

It is interesting that these two quite different implementations of the idea that logical constants are insensitive to changes in cardinal number give so similar results. There are of course other ideas than cardinal insensitivity on which one can base constraints for logical constanthood. Further ideas and results in this direction can be found in van Benthem [1984a; 1983c]. For example, he shows that by slightly weakening *UNIF* one can obtain, in addition to the quantifiers in the square of opposition, **most, not most, least** (i.e. **least** $AB \Leftrightarrow |A \cap B| < |A - B|$), **not least**, and no others, as logical constants. The number tree is an excellent testing ground for experiments in this area.

4.5 *Inference Patterns*

The universal properties of quantifiers we have considered can be seen as *inference schemes for quantified sentences*:

$$\frac{QAB}{QBA} \qquad \frac{\begin{array}{l}QAB\\QBC\end{array}}{QAC} \qquad \frac{\begin{array}{l}QAB\\QAC\end{array}}{QBC} \qquad \text{etc.}$$

(symmetry) (transitivity) (euclidity)

There are also schemes with fixed quantifiers, such as

$$\frac{\begin{array}{l}QAB\\ \textit{all}\ BC\end{array}}{QAC}$$

(*MON*↑).

In 4.1 we answered some questions of the type: which quantifiers satisfy inference scheme S? This is familiar from Aristotle's study of syllogisms, cf. Section 1.1. Aristotle aimed at systematic survey, and he answered the question for *all schemes of a certain form*.

EXAMPLE. Consider schemes with 2 premisses, 1 conclusion (all of the form QXY with distinct X, Y), at most 3 variables, and 1 quantifier symbol. There are 6 possibilities for each formula in a scheme, and hence, up to notional variants (permutations of the variables), $6^3/3! = 36$ possible schemes. Identifying schemes that differ only by the order of the premisses, and deleting the trivially valid schemes whose conclusion is among the premisses, 15 schemas will remain. Then, it can be shown, using Lemma 47 and Theorems 51–53, that for *logical* quantifiers these reduce to *symmetry, transitivity, anti-euclidity*, and the following property which we may call *weak symmetry*:

$$\frac{\begin{array}{l} QAB \\ QBC \end{array}}{QBA}$$

(ignoring unsatisfiable schemes, such as euclidity). Weak symmetry is strictly weaker than symmetry; a number-theoretic characterisation of it can be found in [Westerståhl, 1984].

Thus, there are no other schemes than these (of the present form), and the results of 4.1 (e.g. Corollary 54 and Theorem 56) give us a pretty good idea of which quantifiers satisfy them.

EXAMPLE. *Aristotelian syllogisms*. The schemes are as in the first example, except that there are 3 quantifier symbols, and that one variable (the 'middle term') is required to occur in both premisses but not in the conclusion. Aristotle solved this problem in the special case that quantifiers are taken among **some, all, no, not all**. In the general case of logical quantifiers the solution is of course much more complicated.

The last example indicates that systematic survey of all possible cases is not necessarily an interesting task. In this subsection we shall consider a more specific problem: given the well known inference schemes for basic quantifiers such as **some** and **all**, are these quantifiers *determined* by the schemes, or are the schemes, as it were inadvertently, satisfied by other quantifiers as well?

The logical interest of such questions should be clear. They concern the extent to which the syntactic behaviour of logical constants determine their semantic behaviour. Negative results will tell us that inference rules of a certain type underdetermine semantic interpretation — a familiar situation in logic. Positive results, on the other hand, can be viewed as a kind of *completeness* or *characterisation* theorems.[35]

[35] One analogy is with the usual completeness theorems in logic, relating provability to truth in models. Or, one may think of the extent to which axiomatic characterisations of a relation (say) determine

These questions are also related to deeper issues in the philosophy of language, namely, whether the 'concrete manifestations' of linguistic expressions determine their meaning; cf. post-Wittgensteinian discussions of meaning and use, or Quine's idea on the indeterminacy of translation, or the debate on whether the meaning of logical constants are given by their introduction rules, and more generally on the relation between meaning and proofs (in the context of classical vs. intuitionistic logic; cf. [Prawitz, 1971; Prawitz, 1977; Dummett, 1975]).

Clearly, inference patterns concern the 'concrete' side of language, whereas model theory deals with abstract entities. It would seem that results which relate these two perspectives may be of interest regardless of one's position on the deeper philosophical issues.

A first observation is that the content of our question depends crucially on which kind of inference scheme one allows, i.e. on the choice of *inferential language*. We will look at two such languages here, with quite different properties. But then point is illustrated even more clearly by the following

EXAMPLE. Let the inferential language be predicate logic with the (binary) quantifiers **some** and **all** (this is not essential; we could use $\forall$ and $\exists$ instead). The standard rules for **some**, but with an arbitrary quantifier symbol Q in place of *some*, can be formulated as follows;

$$(1)\quad \frac{\phi(t)\psi(t)}{Qx(\phi(x),\psi(x))} \qquad \frac{\phi(x)\wedge\psi(x)\rightarrow\theta}{Qx(\phi(x),\psi(x))\rightarrow\theta}\ (x \text{ not free in } \theta).$$

$\mathbf{Q}$ *satisfies* a rule of this type if, for each model $\boldsymbol{M}$ and each sequence $\bar{a}$ of individuals from M, if the premisses are true in $(\boldsymbol{M},\bar{a})$ (with Q interpreted as $\mathbf{Q}_M$), then so is the conclusion. But then it is practically *trivial* that

(2) $\mathbf{Q}$ satisfies the rules (1) iff $\mathbf{Q}$ = **some**.

For suppose $\mathbf{Q}$ satisfies (1). Take any M. We must show that $\mathbf{Q}_M AB \Leftrightarrow A\cap B\neq\varnothing$. If $a\in A\cap B$ then P_1x and P_2x are true in $\langle M,A,B,a\rangle$, and hence, by the first rule, so is $Qx(P_1x,P_2x)$, i.e. $\mathbf{Q}_M AB$ holds. If, on the other hand, $A\cap B=\varnothing$, let θ be a logically false sentence and b any element of M. Then $P_1x\wedge P_2x\rightarrow\theta$ is true in $\langle M,A,B,b\rangle$, and thus also $Qx({}_1x,P_2x)\rightarrow\theta$, by the second rule. So $Qx(P_1x,{}_2\,x)$ is false in the model, i.e. $\mathbf{Q}_M AB$ does not hold. (Similar remarks apply to **all**.)

Why does the inferential language of this example trivialise the question of whether the rules characterise the quantifiers? One suggestion might be that rules like (1) are *circular* (in some sense to be specific) as explanations of meaning. In any case, we shall now define two other inferential languages, IL_{syll} and IL_{boole}, for which the problem has non-trivial solutions. These languages have no individual

an intended interpretation (e.g. questions of categoricity). Since the relations in the present case are basic logical constants, a third analogy suggests itself: characterisations of *EL*, such as Lindström's theorem (Section 1.6).

variables, only *set* variables. Most of the inference schemes we have seen so far can be expressed in them. The idea to pose the present charaterisation problem for quantifiers was introduced in [van Benthem, 1984a] and the results on IL_{syll} below are from [van Benthem, 1983c].

DEFINITION of IL_{syll}.

(a) *Syntax: Elementary schemes* in IL_{syll} are of the form QAB or $\neg Q'AB$, where $A, B, \ldots$ are the set variables and $Q, Q', \ldots$ quantifier symbols. A *scheme* in IL_{syll} is either an elementary scheme or has the form

(a) $\phi_1 \wedge \ldots \wedge \phi_n \rightarrow \theta_1 \vee \ldots \vee \theta_k$,

where ϕ_i and θ_j are elementary schemes.

(c) *Semantics*: Suppose ψ is a scheme in IL_{syll} with quantifier symbols among $Q^1, \ldots, Q^m$, and with p set variables. For any quantifiers $\mathbf{Q}^1, \ldots, \mathbf{Q}^m$, a $(\mathbf{Q}^1, \ldots, \mathbf{Q}^m)$-*model* (for ψ)is a model $\mathbf{M} = \langle M, A_0, \ldots, A_{p-1} \rangle$, where Q^i is interpreted as $\mathbf{Q}^i_M$. We say that

$(\mathbf{Q}^1, \ldots, \mathbf{Q}^m)$ *satisfies* the scheme ψ,

if ψ is true (in the obvious sense) in all $(\mathbf{Q}^1, \ldots, \mathbf{Q}^m)$-models. Similarly, $(\mathbf{Q}^1, \ldots, \mathbf{Q}^m)$ satisfies a *set* Ψ of IL_{syll}-schemes if it satisfies each element of Ψ. Finally, the *syllogistic theory* of $(\mathbf{Q}^1, \ldots, \mathbf{Q}^m)$ is

$$Th_{\text{syll}}(\mathbf{Q}^1, \ldots, \mathbf{Q}^m) = \{\psi : (\mathbf{Q}^1, \ldots, \mathbf{Q}^m) \text{ satisfies } \psi\}.$$

This definition just gives more formal versions of notions we have been using all along. For example, all the properties in Table 2 (4.1), *except* antisymmetry and linearity, can be expressed in IL_{syll} (these two would be expressible if we had allowed quantifier *constants* above). That a quantifier $\mathbf{Q}$ satisfies a scheme just means that the scheme expresses a valid inference rule for $\mathbf{Q}$. For example, $\mathbf{Q}$satisfies

$$QAB \rightarrow QBA$$

just in case $\mathbf{Q}$ is symmetric. Note that more than one quantifier symbol may occur in a scheme. For instance, the scheme

$$Q^1AB \wedge Q^2CA \rightarrow Q^1CB$$

is satisfied by the pair (**no**, **all**) (this is the syllogistic inference 'Celarent'; cf. Section 1.1).

DEFINITION of IL_{boole}.

(a) *Syntax*: As for IL_{syll}, except that elementary schemes now have the form QXY or $\neg Q'XY$, where X, Y are (Boolean) combinations of set variables with the symbols $\cap, \cup$, and $-$.

(b) *Semantics*: As before, where the Boolean symbols have their usual meaning.

Examples of schemes in IL_{boole} but not in IL_{syll} are

$$\begin{array}{l} QAB \rightarrow AA\ A \cap B, \\ QAA\cap\ \ B \rightarrow QAB, \\ QAB \rightarrow QA \cap B\ A \cap B, \\ \qquad\quad QA \cap B\ A \cap B \rightarrow QAB; \end{array}$$

the first two together express *CONSERV*, and the other two are (together) equivalent to symmetry.

There is one last

DEFINITION 77. Let Ψ be a set of schemes in IL_{syll} (or IL_{boole}), in the quantifier symbols $Q^1, \ldots, Q^m$. Let $\mathbf{Q}^1, \ldots, \mathbf{Q}^m$ be quantifiers. We say that

$$\Psi \textit{ determines } (\mathbf{Q}^1, \ldots, \mathbf{Q}^m),$$

if (a) $(\mathbf{Q}^1, \ldots, \mathbf{Q}^m)$ satisfies Ψ, and (b) no other sequence of m quantifiers satisfies Ψ. Also, $(\mathbf{Q}^1, \ldots, \mathbf{Q}^m)$ is *determined in* IL_{syll} (IL_{boole}), if some set of schemes in IL_{syll} (IL_{boole}) determines $(\mathbf{Q}^1, \ldots, \mathbf{Q}^m)$.

Note that if $(\mathbf{Q}^1, \ldots, \mathbf{Q}^m)$ is determined in IL_{syll} (IL_{boole}),, it is determined by the set $Th_{\mathrm{syll}}(\mathbf{Q}^1, \ldots, \mathbf{Q}^m)(Th_{\mathrm{boole}}(\mathbf{Q}^1, \ldots, \mathbf{Q}^m))$.

As an example, consider the set consisting of two IL_{syll}-schemes expressing *symmetry* and *quasireflexivity*. **some** satisfies this set, but, by Theorem 64, the set does *not* determine **some**. The obvious question is then whether some larger set determines **some** i.e. whether **some** is determined in IL_{syll}. A negative answer follows from the next theorem.

We assume *FIN* from now on (but see the comments at the end). The quantifiers $\mathbf{some}_n$ and $\mathbf{all}_n$ were defined in Section 4.1.

THEOREM 78 (van Benthem). Th_{syll} *(***some***,* **all***) is satisfied precisely by the pairs* $(\mathbf{some}_n, \mathbf{all}_n)$*, for* $n \geqq 1$.

Thus not even (**some**, **all**) is determined in IL_{syll}. That **some** (or **all**) is not determined follows immediately, since $Th_{\mathrm{syll}}(\mathbf{some}) \subseteq Th_{\mathrm{syll}}(\mathbf{some}, \mathbf{all})$.

This theorem is an immediate consequence of the next two theorems, which give additional information about the pair (**some**, **all**).

THEOREM 79 (van Benthem). $Th_{\mathrm{syll}}(\mathbf{some}, \mathbf{all}) = Th_{\mathrm{syll}}(\mathbf{some}_n, \mathbf{all}_n)$ *for* $n \geqq 1$.

For the next result, let Φ consist of the IL_{syll}-schemes saying that Q^1 is symmetric and quasireflexive and that Q^2 is reflexive and transitive, plus the following schemes:

(4) $Q^1AB \wedge Q^2AC \rightarrow Q^1BC$,

(5) $\neg A^1AA \rightarrow Q^2AB$.

THEOREM 80 (van Benthem). *If* $(\mathbf{Q}^1, \mathbf{Q}^2)$ *satisfies* Φ*, then, for some* $n \geqq 1$, $\mathbf{Q}^1 =$ $\mathbf{some}_n$ *and* $\mathbf{Q}^2 = \mathbf{all}_n$.

The proof uses Theorem 64 and Corollary 60, which tells us that $\mathbf{Q}^1 = \mathbf{some}_m$ and $\mathbf{Q}^2 = \mathbf{all}_k$, for some m, k. It can then be seen that (4) implies that $k \leqq m$, and (5) that $m \leqq k$.

As to the proof of Theorem 79, we shall indicate the basic technique that is used. The first step is reformulation. Note that the negation of a scheme of the form (3) is equivalent to

$$\phi_1 \wedge \ldots \wedge \phi_n \wedge \neg\theta_1 \wedge \ldots \wedge \neg\theta_k,$$

i.e. that *negated schemes* are (equivalent to) conjunctions of elementary schemes. Since

$$\psi \in Th_{\mathrm{syll}}(\mathbf{Q}^1, \ldots, \mathbf{Q}^m) \Leftrightarrow \neg\psi \text{ has no } (\mathbf{Q}^1, \ldots, \mathbf{Q}^m) - \text{model},$$

we are done if any (**some**, **all**)-model for a negated scheme can be transformed into a ($\mathbf{some}_n, \mathbf{all}_n$)-model for the scheme and *vice versa*.

Now let $\boldsymbol{M} = \langle M, A_0, \ldots, A_{p-1}\rangle$ be a (**some**, **all**)-model for $\neg\psi$. Each conjunct in $\neg\psi$ expresses either that a set of the form $A_i \cap A_j$ or $A_i - A_j$ is empty, or that it is non-empty. Each $A_i \cap A_j$ or $A_i - A_j$ can be written uniformly as a union of partition sets of the form $P_s^{\mathbf{M}}$ (cf. Section 1.7). The two types of condition expressed are thus

(a) $x = x_1 + x_2 + \ldots > 0$,

(b) $x = x_1 + x_2 + \ldots = 0$,

where x is the cardinal of $A_i \cap A_j$ (or $A_i - A_j$) and the x_k are the cardinals of the relevant partition sets. Now add $n - 1$ new elements to each *non-empty* partition set. This gives a model $\boldsymbol{M}^+\langle M^+, A_0^+, \ldots, A_{p-1}^+\rangle$, where the conditions (a) and (b) are transformed into

(a$^+$) $x^+ = X_1^+, +x_2^+ + \ldots \geqq n$,

(b$^+$) $x^+ = x_1^+ + X_2^+ + \ldots = 0$.

But then it is easy to check that $\mathbf{M}^+$ is a ($\mathbf{some}_n, \mathbf{all}_n$)-model of $\neg\psi$.

Note that this method does not work if we start with a ($\mathbf{some}_n, \mathbf{all}_n$)-model and want to get a ($\mathbf{some}_{n+1}, \mathbf{all}_{n+1}$)-model, say. For example, with $n = 3$, we may have

$$x = x_1 + x_2 < 3$$

with $x_1 = x_2 = 1$; then adding 1 gives

$$x^+ = x_1^+ + x_2^+ \geqq 4,$$

which means that the schemes of the form $\neg Q^1 A_i A_j$ will not be preserved.

Nevertheless, by an ingenious elaboration of this technique, van Benthem shows that a (**some**$_{n+1}$, **all**$_{n+1}$)-model can in fact always be obtained, and, combining this with yet another construction, he also shows how to obtain a (**some**,**all**)-model from a (**some**$_n$, **all**$_n$)-model.

In view of these negative results about IL_{syll}, it is natural to ask if there is a stronger inferential language where the basic logical constants are determined. Indeed, IL_{boole} is such a language. First observe that in IL_{boole} it is sufficient to look at *one* of the quantifiers **some** and **all**. This follows from the next, easily verified, proposition.

PROPOSITION 81.

(a) $\mathbf{Q}$ *is determined in* IL_{syll} *iff* $\neg\mathbf{Q}$ *is determined in* IL_{syll}.

(b) $\mathbf{Q}$ *is determined in* IL_{boole} *iff* $\mathbf{Q}\neg$ *is determined in* IL_{boole} *iff* $(\mathbf{Q}, \breve{\mathbf{Q}})$ *is determined in* IL_{boole}.

We therefore concentrate on **some**. Let Φ_0 consist of schemes saying that Q is symmetric and quasireflexive, plus the following IL_{boole}-scheme:

(vi) $\neg QAA \wedge \neg QBB \rightarrow \neg QA \cup B\ A \cup B$

THEOREM 82. Φ_0 *determines* **some**.

Proof. Clearly **some** satisfies these schemes. Now suppose $\mathbf{Q}$ is any (logical) quantifier satisfying Φ_0. As before, the first two schemes imply that $\mathbf{Q} = \mathbf{some}_n$ for some $n \geqq 1$. Since $\mathbf{Q}$ satisfies (6), we also have

$$|A| < n \& |B| < n \Rightarrow |A \cup B| < n$$

(for all *sets* A, B). But this means that $n = 1$. ■

Now let us look at the other **some**$_n$ in KL_{boole}. From the last result, Th_{boole} (**some**) $\neq Th_{\mathrm{boole}}$(**some**$_n$) when $n > 1$. The proof technique for IL_{syll} works for IL_{boole} as well — indeed, it works better since conditions on (the cardinal number of) *any* Boolean combinations of $A_0, \ldots, A_{p-1}$ can be expressed there. We thus get a **some**$_n$-model from a **some**-model as before. In fact, even from a **some**$_2$-model we get a **some**$_n$-model with this method: adding $n - 2$ to each non-empty partition set transforms

(a) $x = x_1 + x_2 + \ldots \geqq 2$,

(b) $x = x_1 + x_2 + \ldots < 2$

into

(a)$^+$ $x^+ = x_1^+ + x_2^+ + \ldots \geqq n$,

(b)$^+$ $x^+ = x_1^+ + x_2^+ + \ldots < n$,

since at most one x_i in (b) is non-zero. This gives us

THEOREM 83. $Th_{\mathrm{boole}}(\mathbf{some}_n) \subseteq Th_{\mathrm{boole}}(\mathbf{some}_2) \subseteq Th_{\mathrm{boole}}(\mathbf{some})$, *for* $n > 2$.

No such method works if we start with a $\mathbf{some}_m$-model with $m > 2$, however. This was pointed out by Per Lindström: in fact, we have the

THEOREM 84.

(a) $Th_{\mathrm{boole}}(\mathbf{some}_{n+1}) \not\subseteq Th_{\mathrm{boole}}(\mathbf{some}_n)$, *for* $n \geqq 2$.

(b) On the other hand, if $m \geqq n^2$ *then* $th_{\mathrm{boole}}(\mathbf{some}_m) \subseteq Th_{\mathrm{boole}}(\mathbf{some}_n)$.

Proof.

(a) the case $n = 3$ will give the general idea. Let $\neg\psi$ be a negated scheme in IL_{boole} expressing the conditions

$$
\begin{array}{llll}
x_1 + x_2 + x_3 \geqq k, & x_1 + x + 4 < k, & x_2 + x_4 < k, & x_3 + x_4 < k, \\
x_4 + x_5 + x_6 \geqq k, & x_1 + x_5 < k, & x_2 + x_5 < k, & x_3 + x_5 < k, \\
 & x_1 + x_6 < k, & x_2 + x_6 < k, & x_3 + x_6 < k,
\end{array} \tag{7}
$$

when **Q** is interpreted as $\mathbf{some}_k$ (6 partition sets are needed, so a negated scheme with 3 set variables suffices). First note that for $k = 3$, (7) is satisfied when all the x_i are 1. Thus $\neg\psi$ has a $\mathbf{some}_3$-model. But (7) cannot be true when $k = 4$. For, the first two conditions would give an $x_i (1 \leqq i \leqq 3)$ and an $x_j (4 \leqq j \leqq 6)$ which both are $\geqq 2$, and this contradicts one of the remaining conditions. So $\neg\psi$ has no $\mathbf{some}_4$-model.

(b) Suppose $m \geqq n^2$, and take k such that $(k-1)n \leqq m < kn$. It follows that $n \leqq k$, and hence that $k(n-1) \leqq (k-1)n < m$. Now, given conditions

(a) $x = x_1 + x_2 + \ldots \geqq n$,

(b) $x = x_1 + x_2 + \ldots \leqq n - 1$,

multiply all the x_i by k. Then, $x^+ \geqq m$ in (a)$^+$ and $x^+ < m$ in (B)$^+$; this gives the desired $\mathbf{some}_m$-model.

■

As to the converse inclusions, we have the

THEOREM 85. $Th_{\mathrm{boole}}(\mathbf{some}_n) \not\subseteq Th_{\mathrm{boole}}(\mathbf{some}_m)$, *for* $1 \leqq n < m$.

Proof. Generalising (6), we can write a scheme in ψ in IL_{boole} with $n+1$ set variables which expresses

$$\bigwedge |A_{i_1} \cup \cup \ldots A_{i_n}| < K \Rightarrow |A_1 \cup \ldots \cup A_{n+1}| < k$$

(here the conjunction is taken over all subsets of $\{1, \ldots, n+1\}$ with exactly n elements), when $\mathbf{Q}$ is interpreted as $\mathbf{some}_k$. Then $\mathbf{some}_n$ satisfies ψ. For otherwise, there are sets $A_1, \ldots, A_{n+1}$ such that $|A_1 \cup \ldots \cup A_{n+1}| \geqq n$ and $|A_{i_1} \cup \ldots \cup A_{i_n}| < n$ for $1 \leqq i_1, \ldots i_n \leqq n+1$. it follows that, for all i,

$$A_i \not\subseteq \bigcup_{j \neq i} A_j.$$

So in every A_i there is an element not in the other A_j. But this means that $|A_1 \cup \ldots \cup A_n| \geqq n$, a contradiction.

Now let $m < n$. Choose pairwise disjoint $A_1, \ldots, A_{n+1}$ such that $|A_1| = m - n$ and $|A_i| = 1$ for $1 < i \leqq n+1$. Then, if $1 \leqq i_1, \ldots, i_n \leqq n+1$, the cardinal of $A_{i_1} \cup \ldots \cup A_{i_n}$ is either n or $m-1$, i.e. in both cases $< m$, whereas $|A_1 \cup \ldots \cup A_{n+1}| = m$. So $\mathbf{some}_m$ does not satisfy ψ. ■

Summarising, we find once more that **some** behaves in a significantly different way than $\mathbf{some}_n$ for $n > 1$ (and similarly for **all**):

COROLLARY 86. *Of the quantifiers* $\mathbf{some}_n$*, only* **some** *is determined in* IL_{boole}.

Proof. some is determined, by Theorem 82. Further, if Ψ determines $\mathbf{some}_n$, then, by Theorem 83,

$$\Psi \subseteq Th_{\mathrm{boole}}(\mathbf{some}_n) \subseteq Th_{\mathrm{boole}}(\mathbf{some}).$$

Thus **some** satisfies Ψ, and it follows that $n = 1$. ■

As for the quantifiers satisfying $Th_{\mathrm{boole}}(\mathbf{some}_n)$, it follows from our results here that they are all of the form $\mathbf{some}_k$ with $k \leqq n$, that $\mathbf{some}, \mathbf{some}_2$, and $\mathbf{some}_n$ are always among them, but that $\mathbf{some}_{n-1}$ never is if $n > 3$.

The results in this subsection depend on *FIN*. For IL_{boole}, the proof technique works without *FIN*, but the facts are different. More precisely, with the previous methods one easily proves

THEOREM 87. *For each infinite cardinal* $\kappa, Th_{\mathrm{boole}}(\mathbf{some}) = Th_{\mathrm{boole}}$ *(* **infinitely many***)* = $Th_{\mathrm{boole}}(\mathbf{some}_\kappa)$.

Thus, as one would expect, **some** is not determined in IL_{boole} without *FIN*.

4.6 Local Perspective

Let M be a fixed finite universe, with n elements. We can then study local quantifiers *on* M,with much the same aim as before: of all these quantifiers, which ones are 'realised' in natural language?

Most of our global constraints have local versions. *CONSERV* is the same as before (with M fixed), and so are the monotonicity properties of 3.6 and the relational properties of 4.1. *ISOM* reduces to the local *PERM* (3.3). But one constraint which lacks a local version is *EXT*. As a consequence, results not depending on *EXT* have more or less immediate local versions, but when *EXT* is used, such versions may be harder to get. For example, Theorem 36 on double monotonicity holds locally as well, whereas Corollary 48 on the non-existence of asymmetric quantifiers, which uses *EXT*, fails: $\mathbf{Q}_M AB \Leftrightarrow A = M \& B = \varnothing$ is an asymmetric quantifier on M, satisfying *CONSERV* and *PERM*. Suitably modified versions of Corollary 48 and similar results do exist, however, cf. [Westerståhl, 1983].

One advantage of a local and finite perspective is that the effects of constraints such as *CONSERV* and *PERM* can be assessed in a rather perspicuous way, namely, by the *number* of quantifiers they allow. here are some examples for binary quantifiers on M:

Table 3.

number of quantifiers on M under	no constraints	*CONSERV*	*CONSERV* & *VP*-positivity	*CONSERV* & *MON*↑
no constraints	2^{4^n}	2^{3^n}	2^{2^n}	?
when $n = 2$	65536	512	16	108
PERM	$2^{\binom{n+3}{3}}$	$2^{\binom{n+2}{2}}$	$2^{\binom{n+1}{1}}$	$(n+2)!$
when $n = 2$	1024	64	8	24

There is a simple uniform calculation for the first three entries in both rows of this table (these and other calculations have appeared in [Higginbotham and May, 1981; Keenan and Stavi, 1986; Keenan and Moss, 1985; van Benthem, 1984a; Thijsse, 1983]). Consider a pair (A, B), with $A, B \subseteq M$, as a function f from M to $\{0,1\}^2 : f(x) = (1,1)$ if $x \in A \cap B$, $f(x) = (0,1)$ if $x \in B - A$, etc. There are 4^n such functions and hence 2^{4^n} quantifiers on M. *CONSERV* means that $B - A$

can be assumed to be empty, removing the value $(0, 1)$, and reducing the number of functions to 3^n. By Proposition 30, *CONSERV* + *VP*-positivity means that only the pairs (A, A) need be considered, reducing the number of functions to 2^n.

Under *PERM*, $\mathbf{Q}_M$ is a relation between 4 numbers whose sum is n. To choose such numbers is essentially to put n indistinguishable objects in 4 (distinguished) boxes; there are $\binom{n+3}{3}$ ways to do this, by standard combinatorics. As before, addition of *CONSERV* or *CONSERV* + *VP*-positivity reduces the number of boxes to 3 and 2, respectively.

PERM and *CONSERV* are defined for k-ary quantifiers on $M (k \geqq 2)$, and the above calculations extend straightforwardly to this case: just replace '4' by '2^k' (= the number of partition sets induced by $(A_0, \ldots, A_{k-1})$), '3' by '$2^k - 1$', and '2' by '$2^k - 2$' (in the exponent) in the first two columns of Table 3.

The value $(n + 1)!$ for *Perm* + *CONSERV* + *MON*↑ can be obtained by looking in the *number tree for* M, i.e. the number tree restricted to pairs (x, y) such that $x + y \leqq n$. But the corresponding value without *PERM* is unknown:[36] [Thijsse, 1983] shows that a calculation of this appears to require an explicit calculation of the number of *anti-chains* in $P(M)$; the latter is an unsolved mathematical problem. Thijsse's paper contains several further counting results for quantifiers under various constraints (e.g. the number 108 for the case $|M| = 2$), and so does the paper by Keenan and Moss.

It is rather amazing at first sight that there are 65536 possible quantifiers on a universe with only two elements. The strength of the conservativity universal appears clearly from Table 3, which indicates that counting quantifiers is not just pleasant combinatorics — see the papers by Keenan and Stavi and Keenan and Moss for linguistic applications of such counting results.

Another distinguishing feature of the local perspective on quantifiers is that new *definability* issues arise here. Suppose certain *DET* denotations are *given* in M, and likewise denotations of other expressions: proper names, common nouns, transitive and intransitive verbs, etc. (we may think of a *model* $\boldsymbol{M}$ being given, not just a universe). Suppose further that we have identified certain constructions in natural language which can be interpreted as operations producing new quantifiers

[36] Editors' note. The problem indicated is known as Dedekind's problem: give a nice formula (closed-form expression) for the number of anti-chains in $P(M)$ (or, equivalently, the number of monotone Boolean functions of n variables). As far as I know, the problem is still unsolved. These so-called Dedekind numbers form sequence A000372 in the On-line Encyclopaedia of Integer Sequences, `http://www.research.att.com/~njas/sequences/`: 2, 3, 6, 20, 168, 7581, 7828354, 2414682040998, 56130437228687557907788

The problem also pops up in areas such as tiling and graph colouring. Upper and lower bounds are known (and important for computational purposes), as well as its asymptotic behaviour. The number is well defined and it is rather easy to write a program that calculates the numbers — given sufficient resources. Before 1990 I checked the number for $n = 7$ on a simple PC (one of the values reported in the literature, viz. 2414682040998, turned out to be correct), shortly before 2000 the value for n=8 was calculated. (vide link). FYI: the listed number 108 arises as the product of powers of Dedekind numbers: $2^1 3^2 6^1$, where the exponents are binomial coefficients.

The Editors are grateful to E. Thijsse for this information.

from given denotations. We can then ask which quantifiers can be *generated* from the given denotations by means of these operations. Such generated quantifiers are 'realised' in a definite sense; in fact, if the operations and the starting-point were chosen wisely, one may expect each generated quantifier to be *denoted* by some complex *DET* expression (relative to ***M***).

This approach is pursued in [Keenan and Stavi, 1986]. We will present one of their main results, which shows that *conservativity* is a crucial invariant here. Let $CONSERV_M$ be the class of binary quantifiers on M. Also if K is any class of binary quantifiers on M, let $B(K)$ be the smallest class containing K which is *closed* under conjunction, disjunction, and inner and outer negation. Finally, for each $a \in M$, define the quantifiers $\mathbf{S}_a$ on M by

$$\mathbf{S}_a AB \Leftrightarrow a \in A \cap B.$$

Keenan and Stavi argue that each $\mathbf{S}_a$ can be taken as a basic, initially given quantifier. For, if b is an individual in M who *owns* a and nothing else, i.e. if $P_b = \{a\}$ (cf. Section 2.4.6), then

$$\begin{aligned} \textbf{b's one or more}_M AB &\Leftrightarrow P_b \cap A \subseteq B \& |P_b \cap A| \geqq 1 \\ &\Leftrightarrow \mathbf{S}_a AB. \end{aligned}$$

Note that the $\mathbf{S}_a$ are conservative (but *PERM* fails), and that, to regard them as *given*, we also need each element of M to be given (by proper names or other means), and enough ownership relations to guarantee that for each a in M there is a b in M such that $P_b = \{a\}$. these are not implausible assumptions, and the Boolean operations are natural enough.[37]

THEOREM 88 (Keenan and Stavi). *Suppose* $K \subseteq \text{CONSERV}_M$ *and that* $\mathbf{S}_a \in K$ *for* $a \in M$. *Then* $B(K) = \text{CONSERV}_M$.

Proof. We know from 3.4 that Boolean operations preserve conservativity, so $B(K) \subseteq CONSERV_M$. Now let $\mathbf{Q}$ be any element of $CONSERV_M$. We then have

$$\begin{aligned} \mathbf{Q}AB &\Leftrightarrow \mathbf{Q}A\ A \cap B \\ &\Leftrightarrow \exists X \exists y \subseteq X(\mathbf{Q}XY \wedge X = A \wedge Y = A \cap B) \\ &\Leftrightarrow \bigvee_{\substack{X \subseteq Y \subseteq M \\ \& \mathbf{Q}XY}} (X = A \wedge Y = A \cap B). \end{aligned}$$

Note that the last disjunction is finite. It only remains to show that each disjunct can be generated from the $\mathbf{S}_a$ by Boolean operations. We claim that each disjunct is equivalent to the conjunction of

[37] Cf. [Keenan and Stavi, 1986] for the plausibility of the assumptions. Unlike Keenan and Stavi, I have included inner negation in the closure operations, but this can be avoided at the cost of adding a variant of $\mathbf{S}_a$ (namely, **b's zero or more**, when $P_b = \{a\}$) to the initial quantifiers. In 3.4 I expressed some doubts as to the closure of natural language quantifiers under inner or outer negation. These doubts do not affect Theorem 88, however, for, in the proof, we only apply inner and outer negation to the quantifiers $\mathbf{S}_a$, and, as Keenan and Stavi show, $\neg\mathbf{S}_a$ and $\mathbf{S}_a\neg$ are expressible with familiar *DET*s.

(1) $\bigwedge_{a\in Y} \mathbf{S}_a AB$,

(2) $\bigwedge_{a\in X-y}(\mathbf{S}_a\neg)AB$,

(3) $\bigwedge_{a\in M-X}(\neg\mathbf{S}_a AB = \check{\mathbf{S}}_a AB)$.

For, (1) expresses that $Y \subseteq A\cap B$, (2) that $X-Y \subseteq A-B$, and (3) that $A\cap B \subseteq X$ and $A - B \subseteq X$, and it is easily verified that the conjunction of these expresses that $X = A \wedge Y = A \cap B$. ■

By this theorem, *precisely* the conservative quantifiers on M are generated from certain basic ones by Boolean operations. This lends new significance to the conservativity universal ($U2$). By ($U2$) and the theorem, precisely these quantifiers on M are 'realised', in the sense of being denoted by *DET*s (relative to a model; cf. also note 38).

Note that the complex *DET* expression resulting from the proof of the theorem depends crucially on M. That is, conservative quantifiers, such as **most**, will get *different* 'definitions' on different universes, and there is in general no way of giving a global definition working for all universes. Keenan and Stavi prove a theorem (the 'Ineffability Theorem') to the effect that no fixed *DET* expression, containing symbols for simplex *DET*s, K-place predicates, adjectives, NPs and prepositions, can be made to denote, by varying the interpretation of these symbols, an arbitrary conservative quantifier on an arbitrary universe. The reason is that the number of possible denotations of such expressions grows slower with $|M| = n$ than 2^{3^n}.[38]

5 PROBLEMS AND DIRECTIONS

A basic theme of this paper has been to point to natural language as a source for logical investigation. This theme is by no means limited to quantifiers. Thus, one main direction for further study is *extension to other categories*. Some of the constraints we have studied can be transferred to other categories, and new constraints emerge. A typical trans-categorical constraint is *ISOM*, which has significant effects in most categories. For instance [Westerståhl, 1985a] shows that, for *relations between individuals, ISOM* leaves essentially just Boolean combinations with the *identity relation*, and [van Benthem, 1983b] proves that, for arbitrary *operations on subsets* of the universe, *ISOM* leaves precisely the operations whose values are Boolean combinations of the arguments. For further results in this area, and for a broad assessment of the present approach to logical semantics, the reader is referred to [van Benthem, 1986],which also lists several topics for further research, both in the quantifier area and beyond, complementing the brief suggestions given below.

[38] This makes heavy use of the universal ($U4'$) that simplex *DET*s denote *PERM* quantifiers: $2^{(n+1)(n+2)/2}$ grows slower than w^{3^n}. Without (Ur'), a simplex *DET* symbol could denote any conservative quantifier on any M.

Within the area of quantifiers there is, to begin with, the whole field of the *syntax* of various constructions with *DET*s, and of how to treat them semantically. We have mentioned (Section 2) constructions with *only*, the treatment of definites, of partitives, and of 'there are'-sentences, to take just a few examples. The papers [Keenan and Stavi, 1986; Keenan and Moss, 1985] provide ample evidence that these linguistic questions may be fruitfully pursued from the present model-theoretic perspective.

Another linguistic concern is the search for *universals*. As we have seen, universals can be used as basic theoretical postulates, or they can appear as empirical generalisations, sometimes amenable to explanation by means of other principles. the list of universals in Section 3 was not meant to be complete, and some of the formulations were quite tentative. Further proposals can be found in the papers by Barwise and Cooper and by Keenan and Stavi.

The use of semantic theory to explain linguistic facts, such as the privileged status of certain constants, the restrictions on various syntactic constructions, or the discrepancies between possible and actual interpretations of expressions of a certain category, can most likely be carried a lot further. Recall, for example, the discussion after Table 1 in 3.4. Other similar questions are easily found. Why are there so few simple *VP*-negative quantifiers? Why so few simple *MON* $\downarrow$ ones? Why isn't **not every** a *simple* natural language quantifier (like the other quantifiers in the square of opposition)? Such questions may warrant psychological considerations, but van Benthem's analysis of the 'count complexity' in 4.2 shows that simple model theory may be useful even in this context.

In connection with the last remark, it should be mentioned that van Benthem [1985; 1987a] carries the study of *computational complexity* in semantics much further. He shows (cf. the end of Section 4.3) that the well known complexity hierarchies of automata theory are eminently suitable for classification of quantifiers. Moreover, these investigations carry the promise of a new field of *computational semantics*, which, in addition to questions of logical and mathematical interest, has applications to *language learning* and to mental *processing* of natural language.

On the *logical* side of quantifier theory, many further questions suggest themselves. One natural direction is *generalisation* by weakening the assumptions. For example:

(a) *Drop EXT*. This allows for 'universe-dependent' quantifiers, such as some of the interpretations of *many* in 2.4.3. Some hints on how this admission affects the theory can be found in [Westerståhl, 1983].

(b) *Drop QUANT*. If possessives are allowed, this is a natural move. One can then replace *QUANT(ISOM)* by postulates of *quality*, requiring closure under 'structure-preserving' bijections. Other new constraints can also be formulated for this case, which is studied in [van Benthem, 1983b].

(c) Allow *ternary* quantifiers, or arbitrary n-ary ones ($n \geqq 2$). We did this in Section 3 for the basic concepts, but the corresponding generalisation of the

theory in Section 4 is by no means straight-forward; cf. [Keenan and Moss, 1985].

Dropping *CONSERV*, on the other hand, does not seem fruitful (except for purely logical issues such as definability; cf. Section 4.3). (a)–(c) are not (only) generalisations for their own sake, but linguistically motivated. The next generalisation is more mathematical:

(d) *DROP FIN*. Many of the results using *FIN* can in fact be generalised, as we have noted from time to time. Two apparent exceptions were the results on transitivity, Theorem 56 and Corollary 60 (without *VAR*; cf. Corollary 63). Are there generalisations of these to infinite universes? But perhaps these generalisations lead in the wrong direction. It could be that *FIN*, or some similar constraint, is an essential characteristic of natural language quantification (cf Section 3.8). In any case, the assessment of some minimal model-theoretic means for handling 'natural language infinity' appears to be an interesting task. Some results in this direction can be found in [van Deemter, 1985].

But, even without generalising, the type of logical study conducted in Section 4 can be pursued further. The properties in 4.1 were chosen in a rather conventional way; there may be more interesting *properties of relations* to study. *Definability* questions need not be confined to first-order definability — as we saw in 4.3, *arithmetical definability* is a natural concept in the realm of (logical) quantifiers.

A particularly interesting aspect of definability concerns the *expressive power of natural language*. Various global notions of definability may be used here, e.g. definability *from* given quantifiers. There is also the local definability question mentioned in 4.6: of the possible denotations of expressions of a certain category, which ones are 'generated' in a given model? The conservativity theorem of Keenan and Stavi gives one answer, for *DET* denotations. Perhaps *NP* denotations are even more interesting; this aspect of expressive power is studied in [Keenan and Moss, 1985], where several results on which *NP* denotations are obtainable from quantifiers with certain properties (conservative, logical, *VP*-positive, etc.) are proved.

The study of *inferential languages* from Section 4.5 gives rise to a number of logical questions. This appears to be a recent field, though related to well-known questions on the correlation between a proof-theoretic and a model theoretic perspective on logic.[39] Note that the results of 4.5 depend crucially on our use of *binary* quantifiers instead of unary ones. As for particular questions, one would like to know which quantifiers are determined in these languages. Are *any* (non-trivial) quantifiers determined in IL_{syll}? Are any quantifiers *besides* those in the square of

[39] Zucker [1978] adopts a point of view similar to the present one. There seems to be a connection between his notion of a quantifier being *implicitly definable* and our notion of it being *determined*, even though the settings are different.

opposition determined in IL_{boole}? One can also pose 'finiteness' (compactness) questions, e.g. if **Q** is determined by Ψ, is **Q** by necessity determined by a finite subset of Ψ? This may of course be a trivial question, depending on the answer to the first two. Another compactness question is: if every finite subset of Ψ is satisfied by some quantifier (or sequence of quantifiers), must Ψ itself be satisfiable? Actually, this question can be seen to have a negative answer for IL_{boole}, but the case of IL_{syll} seems open. Other inferential languages could also be considered. In general, one would like to have a better understanding of what is required of a good inferential language. An obvious extension of IL_{syll} and IL_{boole}, however, is to add **some** and **all** as constants. This allows, e.g. monotonicity properties to be expressed in IL_{syll}, and the logical questions are reopened.

In this connection we should also mention an application of the present theory outside the domain of quantifiers: [van Benthem, 1984b] analyses *conditional* sentences *If X then Y* as relations between *sets* $\|X\|$ and $\|Y\|$ (of possible worlds, situations, etc.), i.e. as binary quantifiers, an obtains several interesting results for the logic of conditionals.

Finally, all of the logical questions mentioned so far presuppose the classical model-theoretic framework we have used in this paper. If one wants to treat such linguistically interesting phenomena as *plurals, collective quantification* (as opposed to the *distributive* quantification we have studied; cf. sentences such as *five boys lifted the piano*), or *mass terms* (with new determiners such as *much* or *a little*), this framework has to be extended. From a natural language point of view, such extension seems imperative. For some steps taken in these directions, cf. e.g. [van Benthem, 1983b; Hoeksema, 1983; Link, 1987; Lønning, 1987a; Lønning, 1987b]. An even more radical change would be the switch from the traditional 'static' model theory to a *dynamic* view on interpretation, e.g. along the lines suggested in [Kamp, 1981] or [Barwise and Perry, 1983]. It would be pleasant if the insights gained from the present quantifier perspective were preserved in such a transition. But, however that may be, standard model-theoretic semantics has already, I think, proved unexpectedly useful for a rich theory of quantifiers, and this theory is in turn a fair illustration of the possibilities of a logical study which starts not from mathematics but from natural language.

APPENDIX

A BRANCHING QUANTIFIERS AND NATURAL LANGUAGE

This appendix presents a brief summary of the main issues related to occurrence of branching quantification (Section 1.5) in natural language. A more detailed presentation is given in [Barwise, 1979].

Let us say, somewhat loosely, that a sentence exhibits *proper branching* if its formalisation requires a partially ordered quantifier prefix which is not equivalent to a linear one. There has been some debate over the following question:

(I) *Does proper branching occur in natural languages*?

The debate started with the claim in [Hintikka, 1973] that proper branching occurs in English. Here is the most well known of his examples:

(1) Some relative of each villager and some relative of each townsman hate each other.

The idea is that (1) should be analysed with the Henkin prefix. Arguing that the branching reading of (1) is preferred over linear versions requires a detailed and quite complicated analysis of what we actually mean when using such a sentence, and not all linguists agreed with Hintikka. In [Barwise, 1979], where the main arguments are summarised, it is argued that the most natural logical form of (1) does involve a branching reading, but one which is equivalent to a linear one, so that this branching is not proper. But the answer to (I) does not necessarily depend on sentences like (1). Barwise, who was sympathetic to Hintikka's general claim argued that with other quantifiers that $\forall$ and $\exists$ one can find clearer examples of proper branching. One of his examples was

(2) Most boys in your class and quite a few girls in my class have all dated each other.

It seems that (2) does *not* mean the same as

(3) Most boys in your class have dated quite a few girls in my class

or

(4) Quite a few girls in my class have dated most boys in your class.

The preferred reading of (2) is *stronger* than both of these: it says that there is a set X containing most boys in your class and a set Y containing quite a few girls in my class, such that *any* boy in X and *any* girl in Y have dated each other. Note that X and Y are *independent* of each other. This is a branching reading, which is (provably) not equivalent to any linear sentence in L(**most, quite a few**). We could formalise (2) as

(5)
$$\begin{array}{l} \textit{most } x \textit{ boy}(x) \\ \\ \textit{quite a few } y \textit{ girl}(y) \end{array} \Bigg\} \textit{have dated each other}(x, y).$$

Barwise pointed out that the above truth definition for such sentences gives the desired reading when, as in the present case, both quantifiers are *MON* $\uparrow$, and gave a similar (but different) truth condition for the case when both are *MON* $\downarrow$. He also noted that sentences of this form with one *MON* $\uparrow$ and one *MON* $\downarrow$ quantifier are anomalous.

(6) Few of the boys in my class and most girls in your class have dated each other.

Even though it seems perfectly grammatical, (6) makes no sense, and this may be explained by means of the monotonicity behaviour of the quantifiers involved. Further discussion of the circumstances under which it makes sense to branch two quantifiers can be found in [Westerståhl, 1987].

For another example, van Benthem has noted that we can have proper branching with certain first-order definable quantifiers that are *not* monotone. Consider

(7) Exactly one boy in your class and exactly one girl in my class have dated each other.

The meaning of (7) is clear and unambiguous, and it is easily seen that (7) is not equivalent to any of its 'linear versions' (or to their conjunction). (Note that we are talking about prefixes with *exactly one* here; it is in this sense the branching is proper, even though (7) is clearly equivalent to a (linear) first-order sentence.)

In conclusion, it seems that there are good arguments for an affirmative answer to (I). Then, one may ask:

(II) *What are the consequences for the 'logic of natural language' of the occurrence of proper branching*?

One of the aims of Hintikka's original paper was to use the occurrence of proper branching to give lower bounds of the complexity of this logic. From 1.5 and 1.6 it should be clear that logic with the Henkin quantifier has many affinities with *second-order logic*. In fact, it can be shown that the set of valid sentences with the Henkin quantifiers, or with arbitrary partially ordered prefixes $\forall$ and $\exists$, is recursion-theoretically just as complex as the set of valid second-order sentences, and this is an extremely complicated set. It is tempting to conclude that natural language is at least as complicated. This last inference, however, is not unproblematic. The result about second-order logic depends crucially on the fact that second-order variables vary over *all* subsets (relations) of the universe. In a natural language context, on the other hand, it may be reasonable to *restrict* the range of these variables, and thus to alter the strength of the resulting logic. More on these issues can be found in the chapter by van Benthem and Doets in this Handbook. Some other types of consequences of the occurrence of proper branching are discussed in [Barwise, 1979].

In addition to the principled questions (I) and (II), there is also the more pragmatical:

(III) *Should branching quantification be used more extensively in the analysis of logical and linguistic form*?

Both Hintikka and Barwise suggest that in many cases a branching reading may be preferable regardless of whether the branching is proper or not: the actual *order*

between two (or more) quantifier expressions in a sentence sometimes seems irrelevant, syntactically *and* semantically, and a logical form where these expressions are unordered is then natural. Certain syntactic constructions appear to trigger such branching readings, in particular, conjoined noun phrases with a reciprocal object (*each other*). An even more extensive use of branching is proposed in [van Benthem, 1983a]: he suggests using branching instead of 'substitution' to explain certain well-known scope ambiguities with $\forall$ and $\exists$; cf. also [van Eijck, 1982]. There seem to be a lot of interesting possibilities in this field.

B LOGIC WITH FREE QUANTIFIER VARIABLES

Quantifier symbols have been *constants* in this paper (cf. Section 2.1.3). What happens if they are treated as free variables instead, or, more precisely, as symbols whose interpretation varies with models? From a logical perspective at least, this is a natural question. Some answers are reviewed in this appendix.

To fix ideas, consider a language L_Q, of standard first-order logic with one binary quantifier symbol Q added (for simplicity; we could have added several monadic quantifier symbols, and a fixed (countable) vocabulary of other non-logical symbols. L_Q is a language for logics like $L(\mathbf{most})$, except that this time Q does not denote a fixed quantifier. Instead, a *model* is now a pair $(\boldsymbol{M}, \mathbf{q})$, where $\boldsymbol{M}$ is as before and $\mathbf{q}$ is a binary quantifier on M. Such models are often called *weak models* (since nothing in particular is required of $\mathbf{q}$). *Truth* (satisfaction) in $(\boldsymbol{M}, \mathbf{q})$ is defined in the obvious way, with Q interpreted as $\mathbf{q}$. A *valid* sentence is thus true regardless of the interpretation of Q (and other non-logical symbols). Here is a trivial example:

$$Qx(x \neq x, \psi) \rightarrow (\exists x\phi \vee Qx(\phi, \psi))$$

(where ϕ, ψ only have x free). Are there non-trivially valid sentences in L_Q? This is answered below.

B.1 The Weak Logic

Add to a standard axiomatisation of first order logic the axioms

(1) $Qx(\phi(x), \psi(x)) \leftrightarrow Qy(\phi(y), \psi(y)$
(y free for x in $\phi(x), \psi(x)$)

(2) $\forall x(\phi_1 \leftrightarrow \phi_2) \rightarrow (Qx(\phi_1, \psi) \rightarrow Qx(\phi_2, \psi))$

(3) $\forall x(\phi_1 \leftrightarrow \phi_2) \rightarrow (Qx(\Psi, \phi_1) \rightarrow Qx(\psi, \phi_2))$

(the last two are extensionality axioms for Q).Call this the *weak logic*. Provability (from assumptions) is defined as usual, the deduction theorem holds, and the axiomatisation is obviously sound. The following completeness theorem goes back to [Keisler, 1970]:

THEOREM 89. *If Σ is a consistent set of sentences in the weak logic, then Σ has a weak model.*

Proof.[Outline] A slight extension of the usual Henkin-style proof suffices. Extend Σ to Σ' by witnessing existentially quantified sentences and then to a maximally consistent Γ. Let M consist of the usual equivalence classes $[c]$ of new individual constants, and interpret relation and constant symbols as usual. For each $\psi(x)$ with at most x free, let $\psi(x)^\Gamma = \{[c] \in M : \Gamma \vdash \psi(c)\}$. Then define $\mathbf{q}$ as follows:

$$\mathbf{q}AB \Leftrightarrow \text{ there are } \phi, \psi \text{ such that } \phi^\Gamma =, \psi^\Gamma = B, \text{ and } \Gamma \vdash Qx(\phi, \psi).$$

One then shows that, for all sentences θ,

$$(\boldsymbol{M}, \mathbf{q}) \models \theta \Leftrightarrow \Gamma \vdash \theta$$

by a straight-forward inductive argument, using (1)–(3) and properties of Γ when θ is of the form $Qx(\phi, \psi)$. ■

COROLLARY 90. *The weak logic is complete, compact, and satisfies the downward Löwenheim–Skolem theorem.*

B.2 Axiomatisable Properties of Quantifiers

By the last results, if *all* weak models are allowed, no 'unexpected' new valid sentences appear. However, it may be natural to *restrict* the interpretation of Q to, say, *conservative* quantifiers, or *transitive and reflexive* ones, or *MON* $\uparrow$ ones. Such properties are *second-order*, and hence in general not directly expressible in L_Q. Nevertheless, in many cases the resulting logic is still axiomatisable, by adding the obvious axioms to the weak logic.

Let P be a property of $\mathbf{q}$ expressible by a universal second-order sentence

(4) $\forall X_1, \ldots, \forall X_n \Psi((X_1, \ldots, X_n),$

where the X_i are unary set variables and Ψ is in L_Q (with the X_i acting as predicate symbols). Let the *corresponding set of L_Q-sentences*, Σ_P, consist of the universal closures of all formulas obtained by replacing all occurrences of $X_1, , \ldots, X_n$ in Ψ by L_W-formulas $\phi_1, \ldots, \phi_n$. For example, $\Sigma_{CONSERV}$ and $\Sigma_{MON\uparrow}$ consist, respectively, or universal closures of formulas of the form

$$Qx(\phi, \psi) \leftrightarrow Qx(\phi, \phi \wedge \psi),$$
$$Qx(\phi, \psi) \wedge \forall x(\psi \rightarrow \theta) \rightarrow Qx(\phi, \theta),$$

Let $\mathbf{K}_P$ be the class of models $(\mathbf{M}, \mathbf{q})$ such that $\mathbf{q}$ satisfies P. Clearly,

(5) $(\mathbf{M}, \mathbf{q}) \in \mathbf{K}_P \Rightarrow (\mathbf{M}, \mathbf{q}) \models \Sigma_P$,

but the converse fails in general. To $\mathbf{K}_P$ corresponds a logic, which we write $L(\mathbf{K}_P)$, where truth and validity is as for the weak logic, except that models are restricted to $\mathbf{K}_P$. then is $L(\mathbf{K}_P)$ axiomatised by Σ_P? A sufficient condition is given below.

A subset A of M is called $(\mathbf{M}, \mathbf{q})$-*definable*, if, for some L_Q-formula ψ and some finite sequence $\bar{b}$ of elements of M, $a \in A \Leftrightarrow (\mathbf{M}, \mathbf{q}) \vDash \psi[a, \bar{b}]$. Consider the following property of P:

(*) *If* $(\mathbf{M}, \mathbf{q}) \vDash \Sigma_P$ *then there is a* $\mathbf{q}'$ *satisfying* P *which agrees with* $\mathbf{q}$ *on the* $(\mathbf{M}, \mathbf{q})$-*definable sets.*

we need one more definition: $(\mathbf{M}', \mathbf{q}')$ is an *elementary extension* of $(\mathbf{M}, \mathbf{q})$, in symbols, $(\mathbf{M}, \mathbf{q}) < (\mathbf{M}', \mathbf{q}')$, if $\mathbf{M}'$ is an extension of $\mathbf{M}$ and, for all L_Q formulas ψ and all finite sequences $\bar{b}$ of elements of M, $(\mathbf{M}, \mathbf{q}) \vDash \psi[\bar{b}] \Leftrightarrow (\mathbf{M}', \mathbf{q}') \vDash \psi[\bar{b}]$. Now a straightforward induction proves the

LEMMA 91. *If* $\mathbf{q}$ *and* $\mathbf{q}'$ *agree on the* $(\mathbf{M}, \mathbf{q})$-*definable sets, then* $(\mathbf{M}, \mathbf{q}) < (\mathbf{M}, \mathbf{q}')$.

From this Lemma and Theorem 89 we immediately obtain the

THEOREM 92. *If (*) holds for* P *then each set of* L_Q*-sentences consistent with* Σ_P *in the weak logic has a model in* $\mathbf{K}_P$. *Hence,* $L(\mathbf{K}_P)$ *is complete, compact, and satisfies the Löwenheim–Skolem theorem.*

Instances of this result appear, for example, in [Keisler, 1970; Broesterhuizen, 1975; Sgro, 1977; Makowski and Tulipani, 1977; Barwise, 1978]. To see its utility we consider some examples.

EXAMPLE. Given $(\mathbf{M}, \mathbf{q})$, let M^d be the set of $(\mathbf{M}, \mathbf{q})$-definable subsets of M, and let $\mathbf{Q}^d - \mathbf{Q} \cap (M^d)^2$. If $(\mathbf{M}, \mathbf{q}) \vDash \Sigma_P$ then, since P is universal, $\mathbf{q}^d$ satisfies P *on* M^d. In some cases, $\mathbf{Q}^d$ actually satisfies P on the whole of $P(M)$, i.e. (*) holds with $\mathbf{Q}' = \mathbf{q}^d$. This is true for all the properties of quantifiers in Table 2 (Section 4.1), *except* reflexivity, quasiuniversality and linearity, as is easily checked. So, for example, the logic $L(\mathbf{K}_P)$, where P is the property of being a *strict partial order* (irreflexive and transitive), is axiomatisable.

EXAMPLE. $P =$ *strict linear order*. If $(\mathbf{M}, \mathbf{q}) \vDash \Sigma_P$, let $\mathbf{Q}^*$ be any strict linear order on $P(M) = M^d$, and let $\mathbf{q}' = \mathbf{q}^d + \mathbf{q}^*$ (order type addition). Then $\mathbf{Q}'$ is a strict linear order coinciding with $\mathbf{q}$ on M^d, so $L(\mathbf{K}_P)$ is axiomatisable. As similar construction can be used to show that each of the three properties left over in the preceding example is axiomatisable.

EXAMPLE. $P =$ *MON* $\uparrow$. If $(\mathbf{M}, \mathbf{q}) \vDash \Sigma_{MON\uparrow}$, define $\mathbf{q}'$ by: $\mathbf{q}'AB \Leftrightarrow$ for some $C \in M^d, C \subseteq B$ and $\mathbf{q}AC$. Since $\mathbf{q}$ is *MON* $\uparrow$, $\mathbf{q}'$ agrees with $\mathbf{q}$ on M^d. Also, $\mathbf{q}'$ is *MON* $\uparrow$ (on all subsets of M). Other monotonicity (or continuity) properties can be treated similarly.

EXAMPLE. $P = \mathit{CONSERV}$. If $(\mathbf{M}, \mathbf{q}) \models \Sigma_{\mathit{CONSERV}}$, let $\mathbf{q}'AB \Leftrightarrow \mathbf{q}A\ A \cap B$. Again, the verification that (*) holds is immediate.

EXAMPLE. In the following mathematical example, $\mathbf{q}$ is *unary*, and satisfies P iff $\mathbf{q}^- = P(M) = \mathbf{q}$ is a *proper, non-principal ideal* in $P(M)$, i.e. iff for all $A, B \subseteq M$, (i) $A, B \in \mathbf{q}^- \Rightarrow A \cup B \subseteq \mathbf{q}^-$; (ii) $A \in \mathbf{q}^- \& B \subseteq A \Rightarrow B \in \mathbf{q}^-$; (iii) $M \notin \mathbf{q}^-$; (iv) $\{a\} \in \mathbf{q}^-$ for all $a \in M$. In $L(\mathbf{K}_P), Qx\psi$ can be read 'for many x in the (infinite) universe, ψ'. Now suppose $(\mathbf{M}, \mathbf{q}) \models \Sigma_P$. Then $\mathbf{q}^{d-} = M^d - \mathbf{q}^d$ is a proper, non-principal ideal in M^d. Also, $\mathbf{q}^{d-}$ generates a proper, non-principal ideal $\mathbf{q}'^-$ in $P(M)$: let $A \in \mathbf{q}'^- \Leftrightarrow A \subseteq B_1 \cup \ldots \cup B_n$, for some $B_1, \ldots, B_n \in \mathbf{q}^{d-}$. Then (*) holds for $\mathbf{q}' = P(M) - \mathbf{q}'^-$, so $L(\mathbf{K}_P)$ is axiomatisable. $L(\mathbf{K}_P)$ is studied in [Bruce, 1978], mainly as a mains for obtaining results about the logic $L(\mathbf{Q}_1)$ where $\mathbf{Q}_1$ is the quantifier 'for uncountably many'.

Note that even though axiomatisability comes rather easily in these examples, other properties, such as interpolation, unions of chains, etc. may be much harder an require new methods (cf. [Bruce, 1978]).

C A NON-AXIOMATISABLE PROPERTY

In view of the above examples, one may ask if the property of *quantity* is also axiomatisable. After all, *PERM* is a universal second-order property (with a binary relation variable in addition to the unary set variables), and a corresponding Σ_{PERM} can be found much as before. However, $L(\mathbf{K}_{\mathit{PERM}})$ is a rather strong logic, and *not* axiomatisable. The reason is, roughly, that it can express that two sets have different cardinalities. For example, if $(\mathbf{M}, \mathbf{q}) \in \mathbf{K}_{\mathit{PERM}}$, and $\mathbf{q}MA$ is *not* equivalent to $\mathbf{q}MB$, it follows that either $|A| \neq |B|$ or $|M - A \neq |M - B|$. This is used in the following result, which is due to [Yasuhura, 1969].

THEOREM 93. *Then natural number ordering, $\langle N, <\rangle$, is characterisable in $L(\mathbf{K}_{\mathit{PERM}})$ in the sense that there is an L_Q-sentence θ such that $\langle M, R\rangle$ is isomorphic to $\langle N, <\rangle$ iff, for some $\mathbf{q}$ satisfying PERM, $(\langle M, R\rangle, \mathbf{q}) \models \theta$.*

Proof. Let θ be the conjunction of a sentence saying that $<$ is a linear ordering with immediate successors and a first but not last element, and the sentence

$$\forall x \forall y(\text{`}y \text{ is the successor of } x\text{'} \rightarrow \neg(Qz(z = z, z < x) \leftrightarrow Qz(z = z, z < y))).$$

If $(\langle M, R\rangle, \mathbf{q}) \models \theta$, where $\mathbf{q}$ satisfies *PERM*, it is easy to see that for each $a \in M, |M_a| < |M_{a+1}|$ (where M_a is the set of predecessors to a), and thus that $\langle M, R\rangle$ is isomorphic to $\langle N, <\rangle$. Conversely, if the quantifier $\mathbf{q}$ on N is defined by $\mathbf{q}AB \Leftrightarrow A = N \& |B|$ is even, then *PERM* holds and $(\langle N, <\rangle, \mathbf{q}) \models \theta$. ■

As in Section 1.6, we obtain the

COROLLARY 94. *$L(\mathbf{K}_{\mathit{PERM}})$ is neither complete nor compact.*

Väänänen [1979] extends these results to show that, in terms of *implicit definability* (definability with extra non-logical symbols), $L(\mathbf{K}_{PERM})$ is equivalent to the logic $L(\mathbf{I})$ (cf. 1.6), and that its set of valid sentences is very complicated: it is neither $\Pi_1'^1$ nor $\Sigma_1'^1$ in the analytical hierarchy.

The above theorem and corollary extend, with the same proof, to the logic $L(\mathbf{K}_{PERM+CONSERV})$. They also extend to *logical* quantifiers. To see this, note that in this appendix we have used *local* quantifiers in our models, for which *ISOM* or *EXT* have no immediate meaning. An alternative procedure would be to consider models of the form $(\mathbf{M}, \mathbf{Q})$, where $\mathbf{Q}$ is a global quantifier, and interpret Q as $\mathbf{Q}_M$ on such a model. It is then easy to check that, for each model $(\mathbf{M}, \mathbf{q})$ in $\mathbf{K}_{PERM+CONSERV}$, there is a *logical* quantifier $\mathbf{Q}$ such that $\mathbf{Q}_M = \mathbf{q}$. From this it follows that a sentence is valid in $(L(\mathbf{Q}_{PERM+CONSERV})$ iff it is valid when Q varies over arbitrary logical quantifiers.

Let us remark, finally, that the results of this appendix depend on the fact that the usual universal and existential quantifier constants occur in L_Q. Anapolitanos and Väänänen [1981] show that, if we drop these, and also drop identity, then $L(\mathbf{K}_{PERM})$ becomes axiomatisable; actually it becomes *decidable*.

ACKNOWLEDGEMENT

I am grateful for many helpful comments and suggestions made by several people at various stages of the preparation of this paper, among them, Jens Allwood, Jan van Eijk, Mats Furberg, Franz Guenthner, Björn Haglund, David Israel, Hans Kamp, Ed Keenan, Per Lindström, Barbara Partee, and, in particular, Johan van Benthem.

BIBLIOGRAPHY

[Aczel, 1975] P. Aczel. Quantifiers, games and inductive definitions. In *Proc. of the 3rd Scandinavian Logic Symposium*, pp. 1–14. North-Holland, Amsterdam, 1975.

[Anapolitanos and Väänänen, 1981] Decidability of some logics with free quantifier variables. *Zeit. Math. Logik und Grundl. der Math.*, **27**, 17–22, 1981.

[Barwise, 1978] J. Barwise. Monotone quantifiers and admissible sets. In *Generalised Recursion Theory II*, J. E. Fenstad *et al.*, eds. pp. 1–38. North-Holland, Amsterdam, 1978.

[Barwise, 1979] J. Barwise. On branching quantifiers in English. *J. Phil. Logic*, **8**, 47–80, 1979.

[Barwise and Cooper, 1981] J. Barwise and R. Cooper. Generalised quantifiers and natural language. *Linguistics and Philosophy*, **4**, 159–219, 1981.

[Barwise and Feferman, 1985] J. Barwise and S. Feferman, eds. *Model-Theoretic Logics*, Springer-Verlag, Berlin, 1985.

[Barwise and Perry, 1983] J. Barwise and J. Perry. *Situations and Attitudes*. MIT Press/Bradford, Cambridge, USA, 1983.

[Broesterhuizen, 1975] G. Broesterhuizen. Structures for a logic with additional generalised quantifier. *Colloquium Mathematicum*, **33**, 1–12, 1975.

[Bruce, 1978] K. Bruce. Ideal models and some not so ideal problems in the model theory of $L(Q)$. *J. Symbolic Logic*, **43**, 304–321, 1978.

[Chang and Keisler, 1973] C.C. Chang and K. J. Keisler. *Model Theory*, North-Holland, Amsterdam.

[Church, 1951] A. Church. A formulation of the logic of sense and denotation. In *Structure, Method and Meaning. Essays in Honor of Henry M. Sheffer*, E. Henle *et al.*, eds. pp. 3–24. New York, 1951.

[Cocchiarella, 1975] N. Cocchiarella. A second-order logic of variable-binding operators. *Rep. Math. Logic*, **5**, 9–42, 1975.

[Cooper, 1983] R. Cooper. *Quantification and Syntactic Theory*. D. Reidel, Dordrecht, 1983.

[Cowles, 1981] J. R. Cowles. The Henkin quantifier and real closed fields. *Zeit. Math. Logik und Grundl. der Math.*. **27**, 549–555, 1981.

[Daniels and Freeman, 1978] C. B. Daniels and J. B. Freeman. A logic of generalised quantification. *Rep. Math. Logic*, **10**, 9–42, 1978.

[DeMorgan, 1847] A. DeMorgan. *Formal Logic*, London. Repr. (A. E. Taylor, ed) London, 1926.

[Dummett, 1973] M. Dummett. *Frege: Philosophy of Language*, Duckworth, London, 1973.

[Dummett, 1975] M. Dummett. The philosophical basis of intuitionistic logic. In *Logic Colloquium 73*, H. E. Rose and J. C. Shepherdson, eds. pp. 5–4. North Holland, Amsterdam, 1975.

[Dummett, 1981] M. Dummett. *The Interpretation of Frege's Philosophy*, Duckworth, London, 1981.

[Enderton, 1970] H. B. Enderton. Finite partially-ordered quantifiers. *Zeit. Math. Logik und Grundl. der Math.*, **16**, 393–397, 1970.

[Fenstad *et al.*, 1987] J. E. Fenstad, P.-K. Halvorsen, T. Langholm and J. van Benthem. *Situations, Language and Logic*, D. Reidel, Dordrecht, 1987.

[Flum, 1985] J. Flum. Characterising logics. Chapter III in [Barwise and Feferman, 1985], pp. 77–120.

[Frege, 1892] G. Frege. On concept and object. In *Translations from the Philosophical Writings of Gottlob Frege*, P. Geach and M. Black, eds. Blackwell, Oxford, 1952.

[Frege, 1893] G. Frege. *Grundgesetze der Arithmetik I*, Jena; partial transl. and introduction by M. Furth: *The Basic Laws of Arithmetic*, Univ. Calif. Press, Berkeley, 1964.

[Geach, 1972] P. Geach. A program for syntax. In *Semantics of Natural Language*, D. Davidson and G. Harman, eds. pp. 483–497. D. Reidel, Dordrecht, 1972.

[Goldfarb, 1979] W. Goldfarb. Logic in the twenties: the nature of the quantifier. *J. Symbolic Logic*, **44**, 351–368, 1979.

[Hauschild, 1981] K. Hauschild. Zum Ergleich von Ha"artigquantor und Rescherquantor. *Zeit. Math. Logik und Grundl. der Math.*, **27**, 255–264, 1981.

[Henkin, 1961] L. Henkin. Some remarks on infinitely long formulas. In *Infinistic Methods*, pp. 167–183. , Oxford, 1961.

[Higginbotham and May, 1981] J. Higginbotham and R. May. Questions, quantifiers and crossing. *The Linguistic Review*, **1**, 41–79, 1981.

[Hintikka, 1973] J. Hintikka. Quantifiers vs. quantification theory. *Dialectica*, **27**, 329–358, 1973.

[Hodges, 1983] W. Hodges. Elementary predicate logic. This Handbook, Volume 1.

[Hoeksema, 1983] J. Hoeksema. Plurality and conjunction. In *Studies in Model-theoretic Semantics*, A. ter Meulen, ed. pp. 63–83. Foris, Dordrecht, 1983.

[Kamp, 1981] H. Kamp. A theory of truth and semantic representation. In *Formal Methods in the Study of Language*, J. Groenendijk *et al.*, eds. pp 277–322. Math Centre, Amsterdam, 1981.

[Keenen, 1981] E. L. Keenan. A Boolean approach to semantics. In *Formal Methods in the Study of Language*, J. Groenendijk *et al.*, eds. pp 343–379. Math Centre, Amsterdam, 1981.

[Keenan, 1989] E. L. Keenan. A semantic definition of 'indefinite NP'. 1989.

[Keenan, 1987] E. L. Keenan. Unreducible n-ary quantification in natural language. In *Generalised Quantifiers: Linguistic and Logical Approaches*, P. Gärdenfors, ed. pp. 109–50. D. Reidel, Dordrecht, 1987.

[Keenan and Moss, 1985] E. L. Keenan and L. S. Moss. Generalised quantifiers and the expressive power of natural language. In *Generalised quantifiers in Natural Language*, J. van Benthem and A. ter Meulen, eds. pp. 73–124. Foris, Dordrecht, 1985.

[Keenan and Stavi, 1986] E. L. Keenan and J. Stavi. A semantic characterisation of natural language determiners. *Linguistics and Philosophy*, **9**, 253–326, 1986.

[Keisler, 1970] H. J. Keisler. Logic with the quantifier 'there exists uncountably many'. *Annals of Math Logic*, **1**, 1–93, 1097.

[Kreisel, 1967] G. Kreisel. Informal rigour and completeness proofs. In *Problems in the Philosophy of Mathematics*, I. Lakatos, ed. pp. 138–157. North-Holland, Amsterdam, 1967.

[Lachlan and Krynicki, 1979] A. H. Lachlan and M. Krynicki. On the semantics of the Henkin quantifier. *J. Symbolic Logic*, **44**, 184–200, 1979.

[Ladusaw, 1979] W. Ladusaw. *Polarity Sensitivity and Inherent Scope Relations*, diss., Univ. Texas, Austin, 1979.

[Lindström, 1966] P. Lindström. First order predicate logic with generalised quantifiers. *Theoria*, **32**, 186–195, 1966.

[Lindström, 1969] P. Lindström. On extensions of elementary logic. *Theoria*, **35**, 1–11, 1969.

[Link, 1987] G. Link. Generalised quantifiers and plurals. In *Generalised Quantifiers: Linguistic and Logical Approaches*, P. Gärdenfors, ed. pp. 151–180. D. Reidel, Dordrecht, 1987.

[Lorenzen, 1958] P. Lorenzen. *Formale Logik*, w. de Gruyter, Berlin, 1958.

[Lønning, 1987a] J. T. Lønning. Mass terms and quantification. *Linguistics and Philosophy*, **10**, 1–52, 1987.

[Lønning, 1987b] J. T. Lønning. Collective readings of definite and indefinite noun phrases. In *Generalised Quantifers: Linguistic and Logical Approaches*, P. Gärdenfors, ed. pp. 203–235. D. Reidel, Dordrecht, 1987.

[Łukasiewicz, 1957] J. Łukasiewicz. *Aristotle's Syllogistic*. Clarendon Press, Oxford, 1957.

[Makowski and Tulipani, 1977] J. Makowsky and S. Tulipani. Some model theory for monotone quantifiers. *Archiv f. Math. Logik*, **18**, 115–134, 1977.

[Montague, 1974] R. Montague. *Formal Philosophy*, R. M. Thomason, ed. Yale U. P. New Haven, 1974.

[Mostowski, 1957] A. Mostowski. On a generalisation of quantifiers. *Fund. Math*, **44**, 12–36, 1957.

[Mundici, 1985] D. Mundici. Other quantifiers: an overview. Chapter VI in [Barwise and Feferman, 1985], pp. 211–233.

[Partee, 1984a] B. Partee. Compositionality. In *Varieties of Formal Semantics*, F. Landman and F. Veltman, eds. Foris, Doredrecht, 1984.

[Partee, 1984b] B. Partee. Genitives and 'have', abstract, UMass, Amherst, 1984.

[Patzig, 1959] G. Patzig. *De Aristotelische Syllogistik*, van der Hoeck and Ruprecht, Göttingen; trans. *Aristotle's Theory of the Syllogism*, D. Reidel, Doredrecht, 1968.

[Prawitz, 1971] D. Prawitz. Ideas and results in proof theory. In *Proc. 2nd Scandinavian Logic Symposium*, J. E.Fenstad, ed. pp 235–307. North-Holland, Amsterdam, 1971.

[Prawitz, 1977] D. Prawitz. Meaning and proofs. *Theoria*, **43**, 2–40, 1977.

[Rescher, 1962] N. Rescher. Plurality-quantification, abstract. *J. Symbolic Logic*, **27**, 373–374, 1962.

[Rooth, 1984] M. Rooth. How to get even with domain selection. In *Proc. NELS14*, pp. 377–401, UMas, Amherst, 1984.

[Rooth, 1985] M. Rooth. *Association with Focus*, diss. UMass, Amherst, 1985.

[Russell, 1903] B. Russell. *The Principles of Mathematics*, Allen and Unwin, London, 1903.

[Russell, 1956] B. Russell. *Logic and Knowledge. Essays 1901–1950*, R. C. Marsh, ed. Allen and Unwin (reference to 'Mathematical logic as based on the theory of types', 1908 and 'The philosophy of logical atomism, 1918).

[Sgro, 1977] J. Sgro. Completeness theorems for topological models. *Annals of Math. Logic*, **11**, 173–193, 1977.

[Thijsse, 1983] E. Thijsse. *Laws of Language*, thesis, Rijksunivesiteit Groningen, 1983.

[van Benthem, 1983a] J. van Benthem. Five easy pieces. In *Studies in Model theoretic Semantics*, A. ter Meulen, ed. pp. 1–17. Foris,Dordrecht, 1983.

[van Benthem, 1983b] J. van Benthem. Determiners and logic. *Linguistics and Philosophy*, **6**, 447–478, 1983.

[van Benthem, 1983c] J. van Benthem. A linguistic turn: new directions in logic. In *Proc. 7th Int. Cong. Logic, Methodology and Philosophy of Science*, Salzburg, 1983. R. Barcan Marcus *et al.*, eds. North Holland, Amsterdam, 1986.

[van Benthem, 1984a] J. van Benthem. Questions about quantifiers. *J. Symbolic Logic*, **49**, 443–466, 1984.

[van Benthem, 1984b] J. van Benthem. Foundations of conditional logic. *J. Phil Logic*, **13**, 303–349, 1984.

[van Benthem, 1985] J. van Benthem. Semantic automata. Report No CSLI-85-27, Stanford, 1985.

[van Benthem, 1986] J. van Benthem. *Essays in Logical Semantics*, D. Reidel, Dordrecht, 1985. (Contains among other things, revised version of van Benthem [1983b, c], [1984a, b], [1985].)

[van Benthem, 1987a] J. van Benthem. Towards a computational semantics. In *Generalised Quantifiers, Linguistic and Logical Approaches*, P. Gärdenfors, ed. pp. 31–71. D. Reidel, Dordrecht, 1987.

[van Benthem, 1987b] J. van Benthem. Polyadic quantifiers. To appear in *Linguistics and Philosophy*, 1987.

[van Benthem and Doets, 1983] J. van Benthem and K. Doets. Higher order logic. This Handbook.

[van Deemter, 1985] K. van Deemter. Generalized quantifiers: finite versus infinite. In *Generalised quantifiers in Natural Language*, J. van Benthem and A. ter Meulen, eds. pp. 147–159. Foris, Dordrecht, 1985.

[van Eijck, 1982] J. van Eijck. Discourse representation, anaphora and scope. In *Varieties of S+Formal Semantics*, F. Landman and F. Veltman, eds. Foris, Dordrecht, 1982.

[van Eijck, 1985] J. van Eijck. *Aspects of quantification in natural language*, diss, Rijksuniversiteit, Groningen, 1985.

[Väänänen, 1979] J. Väänänen. Remarks on free quantifier variables. In *Essays on Mathematical and Philosophical Logic*, J. Hintikka *et al.*, eds. pp. 267–272. D. Reidel, Dordrecht, 1979.

[Walkoe, 1970] W. Walkoe, Jr. Finite partially ordered quantification. *J. Symbolic Logic*, **35**, 535–550, 1970.

[Weese, 1981] M. Weese. Decidability with respect to the Härtig and Rescher quantifiers. *Zeitschrift f. Math. Logik und Grundl. der Math*, **27**, 569–576, 1981.

[Westerståhl, 1976] D. Westerståhl. *Some Philosophical Aspects of Abstract Model Theory*, diss., Dept Philosophy, Univ Göteborg, 1976.

[Westerståhl, 1983] D. Westerståhl. On determiners. In *Abstracts from the 7th Int. Congress of Logic, Methodology and Phil of Science*, Vol 2, pp. 223–226, Salzburg, 1983.

[Westerståhl, 1984] D. Westerståhl. Some results on quantifiers. *Notre Dame J. Formal Logic*, **25**, 152–170, 1984.

[Westerståhl, 1985a] D. Westerståhl. Logical constants in quantifier languages. *Linguistics and Philosophy*, **8**, 387–413, 1985.

[Westerståhl, 1985b] D. Westerståhl. Determiners and context sets. In *Generalised quantifiers in Natural Language*, J. van Benthem and A. ter Meulen, eds. pp. 45–71. Foris, Dordrecht, 1985.

[Westerståhl, 1987] D. Westerståhl. Branching generalised quantifiers and natural language. In *Generalised Quantifiers, Linguistic and Logical Approaches*, P. Gärdenfors, ed. pp. 269–298. D. Reidel, Dordrecht, 1987.

[Yasuhura, 1969] M. Yasuhara. The incompleteness of L_P languages. *Fund. Math.*, **66**, 147–152, 1969.

[Zucker, 1978] J. I. Zucker. The adequacy problem for classical logic. *J. Phil. Logic*, **7**, 517–535, 1978.

[Zwarts, 1983] F. Zwarts. Determiners: a relational perspective. In *Studies in Model Theoretic Semantics*, A. ter Meulen, ed. pp. 37–62, Foris, Dordrecht, 1983.

[Zwarts, 1986] F. Zwarts. *Model Theoretic Semantics and Natural Language: the case of modern Dutch*, diss, Nederlands inst, Rijksunivesiteit Groningen, 1986.

Editors' note. The following references have been added to give the reader an overview of recent work

[Altman *et al.*, 2005] A. Altman, Y. Peterzil and Y. Winter. Scope dominance with upward monotone quantifiers. *Journal of Logic, Language and Information*, **14**, 445–55, 2005

[Bach *et al.*, 1995] E. Bach, E. Jelinek, A. Kratzer and B. H. Partee, eds. *Quantification in Natural Languages*, Dordrecht: Kluwer Academic Publishers, 1995.

[Beaver, 1997] D. Beaver. Presupposition. In van Benthem and ter Meulen, 1997, pp. 939–1008.
[Beghelli, 1994] F. Beghelli. Structured quantifiers. In Kanazawa and Piñón, 1994, pp. 119–45.
[Ben-Avi and Winter, 2005] G. Ben-Avi and Y. Winger. Scope dominance with monotone quantifiers over finite domains. *Journal of Logic, Language and Information*, **13**, 385–402, 2005.
[Ben-Shalom, 1994] D. Ben-Shalom. A tree characterization of generalised quantifier reducibility. In Kanazawa and Piñón, 1994, pp. 147–71.
[Boolos, 1998] G. Boolos. *Logic, Logic, and Logic*. Cambridge, MA: Harvard University Press, 1998.
[Chierchia, 1992] G. Chierchia. Anaphora and dynamic binding. *Linguistics and Philosophy*, **15**, 111–84, 1992.
[Cohen, 2001] A. Cohen. Relative readings of 'many', 'often' and generics. *Natural Language Semantics*, **9**, 41–67, 2001.
[Dalrymple *et al.*, 1998] M. Dalrymple, M. Kanazawa, Y. Kim, S. Mchombo and S. Peters. Reciprocal expressions and the concept of reciprocity. *Linguistics and Philosophy*, **21**, 159–210, 1998.
[Feferman, 1999] S. Feferman. Logic, logics and Logicism. *Notre Dame Journal of Formal Logic*, **40**, 31–54, 1999.
[Fernando, 2001] T. Fernando. Conservative generalized quantifiers and presupposition. *Semantics and Linguistic Theory*, **11**: 172–91, 2001.
[Fernando and Kamp, 1996] T. Fernando and H. Kamp. Expecting many. In *Proceedings of the Sixth Conference on Semantics and Linguistic Theory*, T. Galloway and J. Spence, eds. pp. 53–68. Ithaca, NY: CLC Publications, 1996.
[Ginzburg and Sag, 2000] J. Ginzburg and I. Sag. *Interrogative Investigations*. CSLI Lecture Notes 123, Stanford, CA: CSLI Publications, 2000.
[Groenendijk and Stokhof, 1991] J. Groenendijk and M. Stokhof. Dynamic predicate logic. *Linguistics and Philosophy*, **14**, 39–100, 1991.
[Groenendijk and Stokhof, 1997] J. Groenendijk and M. Stokhof. Questions. In van Benthem and ter Meulen, 1997, pp. 1055–1124.
[Hella *et al.*, 1997] L. Hella, J. Väänänen and D. Westerståahl. Definability of polyadic lifts of generalised quantifiers. *Journal of Logic, Language and Information*, **6**, 305–35, 1997.
[Hendriks, 2001] H. Hendriks. Compositionality and model-theoretic interpretation. *Journal of Logic, Language and Information*, **10**, 29–48, 2001.
[Hodges, 1997] W. Hodges. Some strange quantifiers. In *Structures in Logic and Computer Science*, J. Mycielski *et al.*, eds. pp. 51–65. Lecture Notes in Computer Science 1261, Berlin: Springer, 1997.
[Hodges, 2001] W. Hodges. Formal features of compositionality. *Journal of Logic, Language and Information*, **10**, 7–28, 2001.
[Hodges, 2002] W. Hodges. The unexpected usefulness of model theory in semantics. MS, 2002.
[Hodges, 2003] W. Hodges. Composition of meaning (A class at Düsseldorf). Unpublished lecture notes, 2003.
[Kamp and Reyle, 1993] H. Kamp and U. Reyle. *From Discourse to Logic*. Dordrecht: Kluwer, 1993
[Kanazawa, 1994] M. Kanazawa. Weak vs. strong readings of donkey sentences and monotonicity inference in a dynamic setting. *Linguistics and Philosophy*, **17**, 109–58, 1994.
[Kanazawa and Piñón, 1994] M. Kanazawa and C. Piñón, eds. *Dynamics, Polarity and Quantification*, Stanford, CA: CSLI Publications, 1994.
[Keenan, 1992] E. Keenan. Beyond the Frege boundary. *Linguistics and Philosophy*, **15**, 199–221, 1992.
[Keenan, 1993] E. Keenan. Natural language, sortal reducibility, and generalized quantifiers. *Journal of Symbolic Logic*, **58**, 314–24, 1993.
[Keenan, 2000] E. Keenan. Logical objects. In *Logic, Language and Computation: Essays in Honor of Alonzo Church*, C. A. Anderson and M. Zeleny, eds, pp. 151–83. Dordrecht: Kluwer, 2000.
[Keenan, 2003] E. Keenan. The definiteness effect: semantics or pragmatics? *Natural language Semantics*, **11**, 187–216, 2003.
[Keenan, 2005] E. Keenan. Excursions in natural logic. In *Language and Grammar: Studies in Mathematical Linguistics and Natural Language*, C. Casadio, P. Scott and R. Seely, eds., pp. 3–24. Stanford, CA: CSLI Publications, 2005.

[Keenan and Stabler, 2004] E. Keenan and E. Stabler. *Bare Grammar: A Study of Language Invariants*. Stanford, CA: CSLI Publications, 2004.

[Keenan and Westerståahl, 1997] E. Keenan and D. Westerståahl. Generalised quantifiers in linguistics and logic. In van Benthem and ter Meulen, 1997, pp. 837–93.

[Landman, 1989] F. Landman. Groups, *Linguistics and Philosophy*, **12**, 559–605, 723–44, 1989.

[Lappin, 1996a] S. Lappin. Generalized quantifiers, exception sentences and logicality. *Journal of Semantics*, f 13, 197–220, 1996.

[Lappin, 1996b] S. Lappin. The interpretation of ellipsis. In *The Handbook of Semantic Theory*, S¿ Lappin, ed., pp. 145–75. Oxford: Blackwell, 1996.

[Lønning, 1997] J. T. Lønning. Plurals and collectivity. In van Benthem and ter Meulen, 1997, pp. 302–23.

[McGee, 1996] V. McGee. Logical operations. *Journal of Philosophical Logic*, **25**, 567–80, 1996.

[Moltmann, 1995] F. Moltmann. Exception sentences and polyadic quantification. *Linguistics and Philosophy*, **18**, 223–80, 1995.

[Moltmann, 1996] F. Moltmann. Resumptive quantifiers in exception phrases. In *Quantifiers, Deduction and Context*, H. de Swart, M. Kanazawa and C. Piñón, eds., pp. 139–70. Stanford, CA: CSLI Publications, 1996.

[Neale, 1990] S. Neale. *Descriptions*, Cambridge MA: MIT Press, 1990.

[Perry, 2001] J. Perry. *Reference and Reflexivity*, Stanford, CA: CSLI Publications, 2001.

[Peters and Westerståahl, 2002] S. Peters and D. Westerståahl. Does English really have resumptive quantification? In *The Construction of Meaning*, D. Beaver *et al.*, eds, pp. 181–95. Stanford, CA: CSLI Publications, 2002.

[Pustejovsky, 1995] J. Pustejovsky. *The Generative Lexicon*, Cambridge MA, MIT Press, 1995.

[Ranta, 1994] A. Ranta. *Type Theoretical Grammar*. Oxford: Oxford University Press, 1994.

[Recanati, 2002] F. Recanati. Unarticulated constituents. *Linguistics and Philosophy*, **25**, 299–345, 2002.

[Recanati, 2004] F. Recanati. *Literal Meaning*, Cambridge: Cambridge University Press, 2004.

[Sher, 1991] G, Sher. *The Bounds of Logic*, Cambridge, MA: MIT Press, 1991.

[Sher, 1997] G, Sher. Partially-ordered (branching) generalized quantifiers: a general definition. *Journal of Philosophical Logic*, **26**, 1–43, 1997.

[Sperber and Wilson, 1995] D. Sperber and D. Wilson. *Relevance: Communication and Cognition*, Oxford: Oxford University Press, 1995.

[Stanley, 2000] J. Stanley. Context and logical form. *Linguistics and Philosophy*, **23**, 391–434, 2000.

[Stanley, 2002] J. Stanley. Nominal restriction. In *Logical Form and Language*, G. Peters and G. Preyer, eds;, pp. 365–88. Oxford: Oxford University Press, 2002.

[Stanley and Szabo, 2000] J. Stanley and Z. Szabo. On quantifier domain restriction. *Mind and Language*, **15**, 219–61, 2000.

[Sundholm, 1989] G. Sundholm. Constructive generalized quantifiers. *Synthese*, **79**, 1–12, 1989.

[Szabolcsi, 2004] A. Szabolcsi. Positive polarity — negative polarity. *Natural Language and Linguistic Theory*, **22**, 409–52, 2004.

[Thijsse, 1985] E. Thijsse. Counting quantifiers. In em General Quantifiers in Natural Language, J. van Benthem and A. ter Meulen, eds. GRASS 4, Foris, Dordrecht, 1985.

[Väänänen, 1997] J. Väänänen. Unary quantifiers on finite models. *Journal of Logic, Language and Information*, **6**, 275–304, 1997.

[Väänänen, 2001] J. Väänänen. Second-order logic and foundations of mathematics. *Bulletin of Symbolic Logic*, **7**, 504–20, 2001.

[Väänänen, 2002] J. Väänänen. On the semantics of informational independence. *Logic Journal of the IGPL*, **10**, 339–52, 2002.

[Väänänen and Westerståahl, 2002] J. Väänänen and D. Westerståahl. On the expressive power of monotone natural language quantifiers over finite models. *Journal of Philosophical Logic*, **31**, 327–58, 2002.

[Valludví and Engdahl, 1996] E. Valludví and E. Engdah. The linguistic realization of information packaging, *Linguistics*, **34**, 459–519, 1996.

[van Benthem, 1984] J. van Benthem. Questions about quantifiers. *Journal of Symbolic Logic*, **49**, 443–66, 1984. Also in van Benthem, 1986.

[van Benthem, 1986] J. van Benthem. *Essays in Logical Semantics*, Dordrecht: D. Reidel, 1986.

[van Benthem, 1989] J. van Benthem. Polyadic quantifiers. *Linguistics and Philosophy*, **12**, 437–64, 1989.

[van Benthem, 1991] J. van Benthem. *Language in Action*, Amsterdam: North-Holland, 1991; also Boston, MIT Press, 1995.

[van Benthem, 2002] J. van Benthem. Invariance and definability: two faces of logical constants. In *Reflections of the Foundations of Mathematics: Essays in Honor of Sol Feferman*, W. Sieg, R. Sommer and C. Talcott, eds., pp. 426–46. ASL Lecture Notes in Logic, 15, Natick, MA: The Association for Symbolic Logic, 2002.

[van Benthem, 2003] J. van Benthem. Is there still logic in Bolzano's key? In *Bernard Bolzano's Leistungen in Logik, Mathematik und Psysik*, E. Morscher, ed., pp. 11–34. Beiträge zur Bolzano–Forschung, 16, Sankt Augustin: Academia Verlag, 2003.

[van der Does, 1993] J. van der Does. Sums and quantifiers. *Linguistics and Philosophy*, **16**, 509–50, 1993.

[van der Does, 1996] J. van der Does. Quantification and nominal anaphora. In *Proceedings of the Konstanz Workshop "Reference and Anaphoric Relations*, K. von Heusinger and U. Egli, eds;, pp. 27–56, Universität Konstanz, 1996.

[von Fintel, 1993] K. von Fintel. Exceptive constructions. *Natural Language Semantics*, **1**, 123–48, 1993.

[von Fintel, 1994] K. von Fintel. Restrictions on Quantifier Domains. PhD Dissertation, University of Massachusetts, Amherst, 1994.

[Westerståahl, 1991] D. Westerståahl. Relativization of quantifiers in finite models. In *generalized Quantifier Theory and Applications*, j. van der Does and J. van Eijck, eds., pp. 187–205, Amsterdam: ILLC. Also in *idem* (eds.), *Quantifiers: Logic, Models and Computation*, pp. 375–83 Stanford, CA: CSLI Publications.

[Westerståahl, 1994] D. Westerståahl. Iterated quantifiers. In Kanazawa and Piñón, 194, pp. 173–209.

[Westerståahl, 1995] D. Westerståahl. Quantifiers in natural language. A survey of some recent work. In M. Krynicki, M. Mostowski and L. W. Szczerba (eds.), *Quantifiers: Logics, Models and Computation*, pp. 359–408. Kluwer Academic Publishers, 1995.

[Westerståahl, 1996] D. Westerståahl. Self-commuting quantifiers. *Journal of Symbolic Logic*, **61**, 212–24, 1996.

[Westerståahl, 1998] D. Westerståahl. On mathematical proofs of the vacuity of compositionality. *Linguistics and Philosophy*, **21**, 635–43, 1998.

[Westerståahl, 2004] D. Westerståahl. On the compositional extension problem. *Journal of Philosophical Logic*, **33**, 549–82, 2004.

[Zimmermann, 1993] T. E. Zimmermann. Scopeless quantifiers and operators. *Journal of Philosophical Logic*, **22**, 545–61, 1993.

[Zucchi, 1995] S. Zucchi. The ingredients of definiteness and the indefiniteness effect. *Natural Language Semantics*, **3** 33–78, 1995.

[Zwarts, 1998] F. Zwarts. Three types of polarity. In *Plurality and Quantification*, F. Hamma and E. Hinrichs, eds., pp. 177–238. Dordrecht: Kluwer, 1998.

INDEX