Studies in the Philosophy of Sociality 5

Andreas Herzig
Emiliano Lorini *Editors*

# The Cognitive Foundations of Group Attitudes and Social Interaction

Springer

# Studies in the Philosophy of Sociality

Volume 5

More information about this series at http://www.springer.com/series/10961

Andreas Herzig • Emiliano Lorini
Editors

# The Cognitive Foundations of Group Attitudes and Social Interaction

Springer

*Editors*

Andreas Herzig
University of Toulouse
CNRS-IRIT Universite Paul Sabatier
Toulouse Cedex 9, France

Emiliano Lorini
University of Toulouse
CNRS-IRIT Universite Paul Sabatier
Toulouse Cedex 9, France

Studies in the Philosophy of Sociality
ISBN 978-3-319-21731-4    ISBN 978-3-319-21732-1 (eBook)
DOI 10.1007/978-3-319-21732-1

Library of Congress Control Number: 2015953213

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

# Foreword

This volume contains some of the materials presented at the international workshop on ‘The Cognitive Foundations of Group Attitudes and Social Interaction’ that took place in Toulouse on 31 May–1 June 2012. The workshop was one of the major events of the European Network for Social Intelligence (SINTELNET) whose aim was to help build a shared perspective at the intersection of artificial intelligence, the social sciences and humanities, to identify challenges and opportunities for cross-disciplinary collaboration and to provide guidelines for research and policy-making and to kindle partnerships among participants.

The workshop was intended to bring together philosophers, social scientists (economists and psychologists), logicians and computer scientists to discuss about the cognitive foundations of group attitudes and social interaction. It dealt with questions such as:

- What are the relationships between individual attitudes such as beliefs, goals and intentions and group attitudes such as common belief, collective acceptance, joint intentions, group preferences and collective emotions? Can group attitudes be defined from the corresponding individual attitudes, and if so, how? What does it mean that a given group of agents has a collective emotion (e.g. collective guilt, shame, panic)?
- What are the cognitive bases of group identity and group identification (i.e. the fact that an agent identifies himself as members of a given group)? Is group identification reducible to the sharing of ideals and values with the other members of the group? How does group identification influence decision in strategic situations (e.g. team reasoning, I-mode vs. We-mode)?
- What is the role of social emotions such as guilt, shame and envy in social interaction? What are the relationships between social emotions and individual attitudes such as beliefs, goals, intentions, values and ideals? How do these emotions influence decisions in strategic situations?
- What are the relationships between trust and individual attitudes such as beliefs, goals and intentions? Does trust have an affective component? If so, what are the relationships between trust and emotions such as hope and fear, joy and sadness?

- Is game theory sufficient to explain and to model social interaction? Are there concepts that are relevant for explaining and modelling social interaction that are missing in game theory? For example, while the notion of intention has been extensively studied in philosophy of mind and AI, it is not included in the conceptual framework of game theory. Is it important to explain social interaction? If so, how should game theory be extended in order to incorporate this notion?

The present volume contains ten chapters. They offer a broad perspective on different issues and concepts that are situated at the intersection between different disciplines such as cognitive sciences and social psychology, legal theory, logic and artificial intelligence. This includes the concepts of trust and help, the problem of mental representation, the relationship between individual beliefs and group beliefs, the cognitive structure of social emotions and the cognitive bases of norm compliance.

We hope that the material contained in this volume will be useful for improving understanding of the way social concepts and phenomena can be analysed and explained by grounding them on a cognitive foundation.

Toulouse
15 June 2015

Andreas Herzig
Emiliano Lorini

# Contents

# On Help and Interpersonal Control

**Emanuele Bottazzi and Nicolas Troquard**

**Abstract** Help is not much considered in the literature of analytic social philosophy. According to Tuomela (Cooperation – a philosophical study, Springer, 2000), when *a* helps an agent *b* (1) *a* contributes to the achievement of *b*'s goal, and (2) *b* accepts *a*'s contribution to the goal. We take a rather different tack. Our notion of help is unilateral and triggered by an attempt. It is unilateral because we can provide our help to someone without her accepting it. She could be unaware of our actions, or she could be unwilling to receive it. Helping is based on trying because it is agent *b* (supposedly) trying to do something that triggers *a*'s action of help. This is something supported for instance by Warneken and Tomasello's experiments with toddlers (Warneken and Tomasello, Science 311(5765):1301–1303, 2006; Br J Psychol 100:445–471, 2009).

Help is interesting in its own right, but also because it allows us to reconsider the philosophical underpinnings of the essential notion of control in social philosophy. Help is seen here as a kind of weak interpersonal control, where an agent *a*'s agency guides an agent *b*'s agency.

When possible, we evaluate our framework on chosen scenarios taken from the literature in philosophy and psychology. The analysis is driven by a formal, logical approach. In particular, we make use of the modal logics of agency. This assists us in taking sensible philosophical choices, avoiding blatant inconsistencies. Moreover, the resulting formalism has the potential to serve as a computational engine for implementing concrete societies of cooperating autonomous agents.

**Keywords** Social ontology • Help • Control • Agency • Logic

E. Bottazzi • N. Troquard (✉)
LOA-ISTC-CNR, Trento, Italy

LACL, University of Paris-Est Créteil, Créteil, France
e-mail: bottazzi@loa.istc.cnr.it; troquard@loa.istc.cnr.it

A. Herzig, E. Lorini (eds.), *The Cognitive Foundations of Group Attitudes and Social Interaction*, Studies in the Philosophy of Sociality 5,
DOI 10.1007/978-3-319-21732-1_1

# 1 Introduction

Helping behavior manifests itself in virtually every society. In fact, if collective action is an essential constituent of society, it may well be that helping behavior is a prerequisite ingredient of collective action. Instances of help in Human societies are "working as a hospital volunteer", "mailing off a charity donation to help hurricane victims", "cardiopulmonary resuscitation and rescue breathing on someone who has had a heart attack", etc. But helping behavior is commonplace in everyday interactions. It is not just a phenomenon occurring in emergency situations, or when somebody is in real need. There are also more trivial and common ways of helping others. For example "helping someone entering the metro by leaving room for them to get in", "helping a child getting dressed", "helping someone to gather some papers they accidentally dropped in a hallway", etc. In the Encyclopedia of Social Psychology, it is defined as follows:

> Helping behavior is providing aid or benefit to another person. It does not matter what the motivation of the helper is, only that the recipient is assisted. This is distinguished from the more general term prosocial behavior, which can include any cooperative or friendly behavior. It is also distinguished from the more specific term altruistic behavior, which requires that the motivation for assisting others be primarily for the well-being of the other person or even at a cost to oneself. (Kilpatrick 2007, p. 420)

The explanation of the reason for help is best left to social psychology. Although often focused on emergency situations, the study of decisions to help is a typical problem in the discipline. Latané and Darley proposed a decision model of helping (Darley and Latané 1968). Work of classifying helping behavior has also been done. Pearce and Amato (1980) proposed a cognitively-based typology of helping along three dimensions: planned formal versus spontaneous informal; serious versus non serious; and giving or indirect versus doing or direct. Smithson and Amato (1980) extended the classification with one dimension: personal versus anonymous.

If help has been a prominent topic of study in social psychology, the same cannot be said in philosophy. It is true that help is considered in ethics, but the typical questions that are explored there are: Is helping a duty? Are we required to help? Little, instead, has been written in analytic philosophy about what help is. We think that this is a loss, especially in the context of social philosophy. In the last years this stream of studies has been focused on the explanation of complex intertwinings of intentions and actions called *joint actions*. Typical scenarios under investigations are moving a sofa together (Tuomela 2007), painting a house together (Bratman 1992), or preparing a hollandaise sauce together (Searle 1990). All these cases can be readily seen as the sum of some manifestations of help. Therefore we believe that an analysis of help itself may become important to tackle, in further studies, joint actions by means of it.

In our account an archetypical case of help—successful help—occurs when agent $b$ tries to achieve a state of affairs, and $a$ makes sure that, if $b$ is trying to achieve some situation, then that very situation is the case. The contribution someone gives to the realization of that situation can vary. This means, for example, that we help others even if we don't actively intervene into the situation: we see our partner trying to open the door and we help her by just seeing to it that she opens it. If she opens the door without us intervening, we helped her anyway since we, for example already reached the keys in our pocket, ready to open the door for her. As we shall see, this structure is a specialisation of a more general one. Help is a form of control over others' agency. It is a way of monitoring what is going on and if necessary, provide what is needed to accomplish what the helpee is trying to accomplish. It is this preparedness to react as a backup-system that is the relevant part of helping behavior.

We will formalise a general concept of *weak* interpersonal control, a *guidance interpersonal control*, in the modal logics of agency commonly coined "bringing-it-about". See e.g. Kanger (1957/1971, 1972), Pörn (1977), Hilpinen (2001), Elgesem (1993, 1997), Governatori and Rotolo (2005), and Troquard (2014). It is a logic extending propositional logic with one modality $E_i$ for every agent $i$. The formula $E_i\phi$ reads that "agent $i$ brings about that $\phi$", where $\phi$ describes some state of affairs. We will also make use of one modality $A_i$ for every agent $i$, where $A_i\phi$ reads that "agent $i$ tries to bring that $\phi$". A first use of the attempt modality is probably due to Santos and others (Santos and Carmo 1996). In the literature of the "bringing-it-about", influence over agents has been subject to debate. One kind of *strong* interpersonal control—of agent $a$ over agent $b$ for $\phi$—is simply captured by $E_aE_b\phi$.[1] More generally, it is any bringing about or attempt to bring about, by $a$ of some conjunction where at least one conjunct concerns the agency of $b$: $X_a(X_b\phi \wedge \psi)$, where $X_a$ and $X_b$ is some modality of $a$'s and $b$'s agency respectively. In contrast, the pattern of weak interpersonal control will match $X_a(X_b\phi \vee \psi)$ (with $\psi$ typically non-provably equivalent to the logical contradiction $\bot$). By instantiation of our general formalisation of weak interpersonal control, we will be able to discuss a variety of more specific controls, helps, and subjective helps. The logic will allow to express properties pertaining to helping behavior and reason about them rigorously. This will assist us in taking sensible philosophical choices. Moreover, the resulting formalism will have the potential to serve as a computational engine for implementing concrete societies of cooperating autonomous agents.

Control over a certain situation is central in Elgesem's interpretation of the logics of "bringing-it-about" (Elgesem 1993, 1997). Although the language is too abstract to discern all the nuances,[2] its proposed semantics at least offers a modelling guideline of agency in terms of Sommerhoff's model of the goal-directed control that living things possess to achieve their function (Sommerhoff 1969).

One of the main and yet somehow striking points of this kind of logic, is its "static" character. Actions are not considered along their temporal dimension. The notion of change, dynamics and time are abstracted away. Abstraction and

[1] Some authors adopt the maxim *qui facit per alium facit per se* to emphasize that this strong interpersonal control implies full blown agency: $E_aE_b\phi \rightarrow E_a\phi$.

[2] The proposed axiomatisation in Elgesem (1997) indeed requires only simple minimal neighboordhood models that are standard in modal logic. The axiomatisation was refined in Governatori and Rotolo (2005), and proved complete with respect to a class of minimal models.

modularity are the strengths of logic in general. It is because it abstracts away from some details of action makes “bringing-it-about” flexible and easily prone to modular upgrades (Governatori and Rotolo 2005).

It has recently been emphasized that “there is no *one* folk theory of action, in roughly the way there is no one folk tale of Little Red Riding Hood” (Milligram 2010, p. 91). To us, the modal logic of “bringing-it-about” is very useful as a starting tool for the formal analysis of agency, and helping behavior in particular. Since philosophical and logical research on the notion of helping is in its pre-infancy, we believe that abstracting away from some details can be useful to discover at least some of its basic ingredients. As a logic of “doing”, “bringing-it-about” is indeed very apt to capture the essence of cases of *successful* interpersonal control. Successful cases are good starting points to explore tentative interpersonal control and helping behavior, as well as more “epistemic” cases. These are cases of being helpful but possibly ineffective. Trying replaces doing, and imperfect information brings in interesting troubles. Hence, the strength of the logic putatively lies in its very abstractness, as one can abstract away from distracting phenomena and still incorporate them later ahead in the analysis.

## 2 Guidance Interpersonal Control

Logics of agency, and logics of “bringing-it-about” specifically, are the logics of the modalities $E_x$ where $x$ is an acting entity, and $E_x\phi$ reads “$x$ brings about $\phi$”, or “$x$ sees to it that $\phi$”. This tradition in logics of action comes from the observation that action is better explained by what it brings about. It is a particularly adequate view for *ex post acto* reasoning, and thus for discovering whether an acting entity is responsible at the moment of the achievement of an action. In a linguistic analysis of action sentences, Belnap and others (Belnap and Perloff 1988; Belnap et al. 2001) adopt the *paraphrase thesis*: a sentence $\phi$ is agentive for some acting entity $x$ if it can be rephrased as $x$ sees to it that $\phi$. Under this assumption, all actions can be captured with the abstract modality. It is regarded as an umbrella concept for direct or indirect actions, performed to achieve a goal, maintaining one, or refraining from one.

The philosophy that grounds the logic was carefully discussed by Elgesem in Elgesem (1993). Suggested to him by Pörn, Elgesem borrows from theoretical neuroscientist Sommerhoff (1969) the idea that agency is the actual bringing about of a goal towards which an activity is directed. Elgesem’s analysis leans also on Frankfurt (1988, Chap. 6) according to whom, the pertinent aspect of agency is the manifestation of the agent’s guidance towards a goal. Sommerhoff’s goals are not necessarily goals proper, and instead are *telos* of an activity, that is, its terminus or end. This means that the notion of bringing about may refer also to non-intentional actions (Hilpinen 1997; Governatori and Rotolo 2005) that have a final end anyway, related for example, to mere instinct.[3]

[3] The main source is *Nicomachean Ethics* III. For a recent review on Aristotle’s voluntariness of action see Meyer (2006).

In Aristotelian terms, action (*praxis*) and production (*poiesis*) have, as their object, the contingent, that which can be otherwise (*to endechomenon allos echein*[4]). It is an important issue for whom is working in modal logic of agency. It was at the core of the discussions in early work such as Kanger (1957/1971, 1972), and Pörn (1977), and in more recent examinations (Hilpinen 2001). The crux of the issue is to capture the idea, in the semantics, that what the agent brings about has to be avoidable. In philosophy this is traditionally seen as *control*. To exercise control, possibilities have to be open to the agent. And this amounts to say that to bring about a state of affair $\phi$ is to exercise a control on $\phi$. In Kanger (1957/1971, 1972) this is linked to what is called *negative* condition of agency, that can be termed *counterfactual condition*, saying that if the agent had not acted the way she did, $\phi$ *would have not* been obtained (Hilpinen 2001). The exact nature of this negative condition has been open to debate ever since. To mention only an eminent proposal, according to Pörn (1977) this condition has to be weakened to the point that if the agent had not acted the way she did, $\phi$ *might have not* been the case.

With respect to that, Elgesem (1997) makes an interesting point, holding that even this weaker negative condition is too strong. One can imagine cases where $\phi$ is the case, independently of what the agent does, but where it is still the case that he brings it about that $\phi$:

> Consider this example. My one-year-old boy is in the process of learning to eat by himself. Sometimes he succeeds in getting the food into his mouth with the spoon, and sometimes not. Suppose he succeeds at some point during the meal, i.e. he brings it about that he has food in his mouth. During the whole of this meal, I am watching him to make sure that he gets fed. So if he does not succeed in getting the food into his mouth, I put the food into his mouth anyway. In this situation, it seems to be the case that there is no relevant alternative where it is not true that he gets food in his mouth. Now, in the case where he hits his mouth with the spoon, it must be right to say that he brings it about that he has food in his mouth. This is the case despite the existence of a reliable back-up system which guarantees that the goal is satisfied in any case.

In the context of the present study, the baby scenario is noteworthy for two reasons. The first one is that it suggests that we should consider a different notion of control. According to John Martin Fischer (1994, 2012), we can isolate at least two kinds of control: *regulative* control and *guidance* control. Regulative control is conceptually linked with the negative condition, because it requires freedom to choose and do otherwise. The notion of guidance control stems from Harry Frankfurt (1969) and gives a better account of Elgesem's stance, because it does not require to consider necessarily that something can be otherwise. Guidance "is determined by characteristics of the actual sequence issuing in one's choice" (O'Connor 2014). Fischer, in order to illustrate this notion, proposes the example of driving a car. I have regulative control of the car if, given the fact that I wish to make a right turn, the car, as a result, moves to the right, but given the perfect condition of the car, I could have decided to make it turn to the left. Instead, suppose that the car is not in perfect condition, but has a quite peculiar malfunctioning such

[4] *NE* IV 5, 1140b 27.

that, if I steer to the right it does it perfectly, but if I try and steer to the left, the car goes to the right, too. Suppose now that I actually steer the wheel to the right (the direction that does not display the malfunctioning). In this case I have guidance control of the car.

The second reason to find interest in Elgesem's scenario is that here, we are not simply dealing with agency, but with *interpersonal* agency. An interpersonal action, as justly observed by Seumas Miller (2002), is an action that is *interdependent* with the action of some other agent, or is otherwise directed to an agent. In the case of the baby, the controlling agency is not just putting some food in a cavity, it is making sure that if the baby tries to put some food in his mouth, the food *is* in his mouth. The result may be realized with the contribution of his father, or by the baby actually feeding himself. We can call this *weak* or *guidance interpersonal control*, having two main components: the controlling agency and the controlled agency. The controlling agency is, in Elgesem's scenario, the father bringing about that *the baby is fed if the baby tries to be fed*.

In guidance interpersonal control, the controlling agency does not have "to go the way of" the controlled agency, though. Take for instance a case of counter-action. Imagine a rush-hour traffic scenario, and in particular these two cars side by side on two different lanes. When the driver in the car on the right lane (say Dr. R) tries to slot his car into the left lane, the driver in the car on the left lane (say Dr. L) will accelerate ever so slightly to prevent it. The controlled agency is Dr. R's trying to have his car on the left lane. The controlling agency is here Dr. L bringing about that *Dr. R's car is not on the left lane if Dr. R tries to slot his car into the left lane*.

The notion of control that we want to highlight is a form of weak control indeed. Not only because interpersonal control is not regulative, but also because the control we are interested in is not a coercion of the controlled agent. It is not a form of constraining the agent into an unavoidable action, and it is not a form of mind control. It is control over a situation in which another agent is actively involved, and has an autonomously acquired volition. In Elgesem's son example the baby is not force-fed, but simply fed by his father when he tries and fails to do it by himself.

*Trying*, or attempt will become crucial in our work here. It has been analyzed in the philosophical literature, considered as a common feature of human actions and often linked to volition. (See Hornsby (1980) and O'Shaughnessy (1980). See also Lorini and Herzig (2008) for a review of the philosophical literature from a logical standpoint.) Trying clearly differs from effective agency. One can bring about something without even wanting it, for example by mistake. The *telos* of some bringing about is the final end of the action. It is in a way, where the action is directed, and it is not necessarily linked with volitions. When someone brings about that $\phi$, we can say that $\phi$ is true. On the other end, when we consider the notion of trying, volition enters into the picture and from the fact that someone tries $\phi$ we cannot infer that $\phi$ is true. As highlighted also in the recent literature (Hornsby 2010), one of the non obvious points related to trying is assessing whether someone has tried to do something whenever she has *succeeded* in doing it. We finally take no stance on the issue: we do not think that a bringing about implies an attempt, and we do not think that an attempt implies a bringing about. These principles will shine by their absence in the logic presented in the next section.

# 3 Logical Aspects of Guidance Interpersonal Control

## 3.1 The Logic of "Bringing-It-About" as a Starting Point

In this paper, we will use the logics of bringing-it-about (BIAT). It has been studied over several decades in philosophy of action, law, and in multi-agent systems (Kanger and Kanger 1966; Pörn 1977; Lindahl 1977; Elgesem 1993, 1997; Santos and Carmo 1996; Santos et al. 1997; Royakkers 2000; Gelati et al. 2004; Troquard 2014). It is the logic of the modality $E_i$, where $i$ is an agent, and $E_i\phi$ reads "$i$ brings about that $\phi$". Following Santos et al. (1997), we will also integrate one modality $A_i$ for every agent $i$, and $A_i\phi$ reads "$i$ tries to bring about $\phi$".

We have laid out the main conceptual foundations of these operators in Sect. 2. We are now going concentrate on the formal features of their logic.

Throughout the paper, we will assume a finite set of agents $\mathsf{Agt}$ and an enumerable set of atomic propositions $\mathsf{Atm}$. The language of BIAT extends the language of propositional logic over $\mathsf{Atm}$, with one operator $E_i$ and one operator $A_i$ for every agent $i \in \mathsf{Agt}$.

The language $L$ is defined by the following grammar:

$$\phi ::= p \mid \neg\phi \mid \phi \wedge \phi \mid E_i\phi \mid A_i\phi$$

where $p \in \mathsf{Atm}$, and $i \in \mathsf{Agt}$.

The fundamental principles (axioms schemes and rules and inference) of BIAT (where $i$ is an individual agent) are[5]:

(prop) $\vdash \phi$, when $\phi$ is a tautology of classical propositional logic
(notaut) $\vdash \neg E_i \top$
(success) $\vdash E_i\phi \rightarrow \phi$
(aggreg) $\vdash E_i\phi \wedge E_i\psi \rightarrow E_i(\phi \wedge \psi)$
(ree) if $\vdash \phi \leftrightarrow \psi$ then $\vdash E_i\phi \leftrightarrow E_i\psi$
(rea) if $\vdash \phi \leftrightarrow \psi$ then $\vdash A_i\phi \leftrightarrow A_i\psi$

BIAT extends propositional classical logic (prop). An acting entity never exercises control towards a tautology (notaut). Agency is an achievement, that is, the culmination of a successful action (success). Agency aggregates (aggreg). Agency and attempts are closed under provably equivalent formulas (ree) and (rea).

We keep the logic of $A_i$ very minimal. In particular, we do not take for granted that every actual agency requires an attempt. That is, we do not integrate $E_i\phi \rightarrow A_i\phi$ in the previous Hilbert system.

[5] For any formula $\phi$, the notation $\vdash \phi$ means that $\phi$ is provable within the logic. It is a theorem of the logic. That is, it is an axiom or a formula that can be deduced from the axioms and rules of inferences.

It is important to note that neither $E_i\phi \to E_i(\phi \vee \psi)$ nor $A_i\phi \to A_i(\phi \vee \psi)$ are derivable. They would indicate that agency and attempt are monotone modalities. We do not want that a bringing about that the letter is posted necessarily implies a bringing about that the letter is posted or the letter is burnt. In fact, adding the former formula to the axiomatization would yield an inconsistent theory. (In classical logic, it is incompatible with (notaut).) The logic of bringing it about is a weaker version of the achievement stit and of the deliberative stit in Belnap et al. (2001). It is different from the Chellas' stit (Horty 2001) which does admit the monotony of agency.

**Strong interpersonal control.** One typical kind of *strong* interpersonal control occurs when agent $a$ brings about that an agent $b$ brings about that $\phi$. It is captured by $E_aE_b\phi$

More generally, a strong interpersonal control is any bringing about or attempt to bring about, by $a$ of some conjunction where at least one conjunct concerns the agency of $b$. That is, where $X_a$ and $X_b$ is some modality of concerning $a$'s and $b$'s agency respectively:

$$X_a(X_b\phi \wedge \psi)$$

**Decidability.** The decidability of BIAT is important for its practical application in reasoning about social situations and procedures. The proof is a simple adaptation of the result obtained in Troquard (2014).

**Proposition 1.** *Let a formula $\phi$ in the language of BIAT. The problem of deciding whether $\vdash \phi$ is decidable.*

This means that we can algorithmically decide of the validity of any property expressed in the language of BIAT. There is a procedure that one can mechanically follow, that will eventually provide the right answer to the question "is the formula $\phi$ valid?", for every formula $\phi$. There is a practical limitation in that the time complexity may grow exponentially with the size of the formula one wishes to automatically analyze. However, the task can be performed without an exponential blowup in space complexity.

The base logic is decidable but we do not claim so for the logics obtained by extending the above Hilbert system as suggested in the remaining of this paper. The problem for each single extension would require to be addressed individually.

### *3.2 The General Form of Guidance Interpersonal Control*

We have seen in the previous sections that interpersonal control involves two interweaving actions: the controlled agency performed by a controlled agent $b$, and the controlling agency, performed by a controlling agent $a$. The latter capitalizes on the former to achieve some state of affairs, say $\gamma$. In order to logically characterize interpersonal control, we introduce three instrumental modalities, intended to capture *agentive modes*. We list them below along with a rough description of their purpose:

1. $X_1^{ab}$: used to capture the *mode of the controlling agency*;
2. $X_2^{ab}$: used to capture the *mode of the content of the controlling agency*;
3. $X_3^{ab}$: used to capture the *mode of the controlled agency.*

To reflect that the agentive modes are indeed modalities, we only need to assume that the obey the rule of equivalents:

$$\begin{array}{l}(\text{rex1}) \text{ if} \vdash \phi \leftrightarrow \psi \text{ then} \vdash X_1^{ab}\phi \leftrightarrow X_1^{ab}\psi \\ (\text{rex2}) \text{ if} \vdash \phi \leftrightarrow \psi \text{ then} \vdash X_2^{ab}\phi \leftrightarrow X_2^{ab}\psi \\ (\text{rex3}) \text{ if} \vdash \phi \leftrightarrow \psi \text{ then} \vdash X_3^{ab}\phi \leftrightarrow X_3^{ab}\psi\end{array}$$

The modalities must be expressible in the language but can take many forms. Example of modalities $X$ that can be defined are $X\phi = E_i\phi$, $X\phi = E_i\neg A_j\phi$, $X\phi = E_j\neg A_i\phi$, $X\phi = A_j\phi \wedge E_i\neg\phi$, etc. For each example, it can indeed be readily checked that if $\vdash \phi \leftrightarrow \psi$ then $\vdash X\phi \leftrightarrow X\psi$. Despite this generality, we will frame more specifically the modalities intended to be used below.

*Remark 1.* Instead of giving the modalities $X_1^{ab}$, $X_2^{ab}$, and $X_3^{ab}$ a definition proper, we will use an axiomatic definition. For instance, instead of defining $X\phi = E_i\phi$, we would adopt $\vdash X\phi \leftrightarrow E_i\phi$ as an axiom. In such a way, we can flexibly provide partial, underspecified definitions of modalities. A weaker version of the previous example could be given as $\vdash X\phi \rightarrow E_i\phi$.

The definition of a *guidance interpersonal control of agent a over agent b for* $\gamma$ is then as follows:

$$\mathsf{GIC}(X_1^{ab}, X_2^{ab}, X_3^{ab}, \gamma) \overset{\text{def}}{=} X_1^{ab}(X_2^{ab}\gamma \rightarrow \gamma) \wedge X_3^{ab}\gamma$$

It is a general account for a situation, or state of affairs, describing $a$'s controlling agency over $b$'s agency, to obtain $\gamma$.

It is now better to progressively deconstruct $\mathsf{GIC}(X_1^{ab}, X_2^{ab}, X_3^{ab}, \gamma)$. It is the general form of guidance interpersonal control and is a conjunction of two distinct states of affairs pertaining to some kind of agency:

- $X_1^{ab}(X_2^{ab}\gamma \rightarrow \gamma)$ is the controlling agency of the guidance interpersonal control;
- $X_3^{ab}\gamma$ is the controlled agency of the guidance interpersonal control.

In the controlled agency:

- $X_3^{ab}$ is the agentive mode of the controlled agency.

In the controlling agency:

- $X_1^{ab}$ is the agentive mode of the controlling agency;
- $X_2^{ab}\gamma \rightarrow \gamma$ is the content of the controlling agency;
- $X_2^{ab}$ is the agentive mode of the content of the controlling agency.

Typically then, we will think of the modalities reflecting more specific modes than suggested before. The modality $X_1^{ab}$ would reflect some agentive mode pertaining to $a$. To commit the definition to a more definite flavor of *control* of $a$, we will consider that $X_1^{ab}$ is either $A_a$ or $E_a$. Practically, it means that we will only consider such instantiations in this paper. Agent $a$'s control is over $b$'s agency. So in the instantiations of guidance interpersonal control considered in this paper, the modalities $X_2^{ab}$ and $X_3^{ab}$ will always reflect some agentive mode pertaining to $b$. The main idea is that (i) the controlled agency indeed reflects $b$'s agency, (ii) the controlling agency indeed reflects $a$'s agency, and (iii) the content of the controlling agency partly reflects $b$'s agency.

**The rush-hour traffic scenario.** Remember Dr. R trying to slot his car into the left lane, and Dr. L making sure that it does not happen if he does try. Take $\mathsf{left}$ to mean that Dr. R's car is on the left lane. The guidance interpersonal control at play in the scenario can be instantiated as follows:

$$\mathsf{GIC}(E_{DrL}, A_{DrR}\neg, A_{DrR}\neg, \neg\mathsf{left})$$

which translates into:

$$E_{DrL}(A_{DrR}\mathsf{left} \rightarrow \neg\mathsf{left}) \wedge A_{DrR}\mathsf{left}$$

## 3.3 Some Formal Properties of Interpersonal Control

As a general definition, our formal account of guidance interpersonal control (of $a$ over $b$) can be instantiated to specific cases by simply identifying the three agentive modes to a particular modality expressible in the logic of BIAT. We can then use the formal tools provided by the logic to rigorously define a terminology pertaining to the properties of interpersonal control. We begin with a few simple qualities.

An interpersonal control is *well-situated* when the agentive mode of the controlled agency coincides with the agentive mode of the content of the controlling agency.

**Definition 1 (WS).** For any $\psi$, a guidance interpersonal control $\mathsf{GIC}(X_1^{ab}, X_2^{ab}, X_3^{ab}, \psi)$ is *well-situated* when

$$\vdash X_3^{ab}\psi \leftrightarrow X_2^{ab}\psi$$

Intuitively, well-situatedness is a good property for $a$'s controlling agency. Indeed, in a well-situated controlled agency (for $\gamma$), $\gamma$ is true iff the content of the controlled agency is true.[6] With the right mode of controlling agency, $a$'s agency can then be the least effort for $\gamma$.

[6] We have $\vdash (\mathsf{GIC}(X_1^{ab}, X_2^{ab}, X_3^{ab}, \gamma) \wedge (X_3^{ab}\gamma \leftrightarrow X_2^{ab}\gamma)) \rightarrow ((X_2^{ab}\gamma \rightarrow \gamma) \leftrightarrow \gamma)$.

What more sure way to have the content of a well-situated controlled agency true, and hence $\gamma$, than to effectively bring it about? An interpersonal control is *effective* when the agentive mode of the controlling agency coincides with a bringing about of agent $a$.

**Definition 2 (EF).** For any $\psi$, a guidance interpersonal control $\mathsf{GIC}(X_1^{ab}, X_2^{ab}, X_3^{ab}, \psi)$ is *effective* when

$$\vdash X_1^{ab}\psi \leftrightarrow E_a\psi$$

Uncertain, unskilled, or hazardous controlling agency by $a$ would remain a worthwhile effort. We say that an interpersonal control is *tentative* when the agentive mode of the controlling agency coincides with an attempt of agent $a$.

**Definition 3 (TE).** For any $\psi$, a guidance interpersonal control $\mathsf{GIC}(X_1^{ab}, X_2^{ab}, X_3^{ab}, \psi)$ is *tentative* when

$$\vdash X_1^{ab}\psi \leftrightarrow A_a\psi$$

Remember that we framed earlier $X_1^{ab}$ to be identified either as $A_a$ or as $E_a$. No other mode of controlling agency will be considered here. In this context, Definitions 2 and 3 offer a clear dichotomy of guidance interpersonal control: (i) effective control, which is actual and successful,[7] (ii) tentative control, which is an uncertain, possible control.

These properties of interpersonal control are presented as provable logical formulas in the language of BIAT. Methodologically, it means that to design a logic of guidance interpersonal control, it suffices to combine into a Hilbert system:

1. the axiomatic system of BIAT presented on Page 7;
2. the principles (rex1), (rex2), and (rex3);
3. the definition of $\mathsf{GIC}(X_1^{ab}, X_2^{ab}, X_3^{ab}, \gamma)$;
4. a set of properties of guidance interpersonal control.

In Sect. 5, we will address a few properties of interpersonal control that are more specific to the notion of help, and we propose a few simple theorems to exemplify the kind of reasoning that is enabled by our formal proposal.

## 4 Help

Help is a form of *weak, interpersonal, guidance control of agent a over agent b for some state of affairs $\phi$*. It goes, so to say, in the way of the helpee's trying or attempting. The controlling agency here is for the sake of the controlled one. In

[7] Successful in virtue of axiom (success).

order to provide help, some control over the situation in which the helpee is trying to achieve some $\phi$ is needed. The baby example provided in Sect. 2 is not just an example of interpersonal guidance control, but it is also an example of *help*. The father is helping his child to get fed, and this does not mean that all the time he is materially putting the food in his mouth. The father is exercising guidance control of a disjunctive state of affairs involving the child tryings to get fed. If the child is able to get fed by himself, there is no need to intervene. On the other hand, if the child tries but fails, then the father's agency will have him intervene in guiding the food in the mouth of the baby. In contrast, the rush-hour traffic example (Sect. 2, and end of Sect. 3.2), although a weak interpersonal guidance control, certainly is not a helping behavior.

As we said the notion of help is not much studied in philosophy in its own right. A notable exception, Raimo Tuomela (2000) endorses that helping is in essence *a* adopting *b*'s goal and *b* accepting it. It is then a special case of asymmetric cooperative activity. If *b* has much to carry and *a* has no load, *a* may offer to help *b* to carry some of *b*'s bags. In thus helping *b*, *a* engages in a cooperative activity with *b* and *b* accepts *a*'s help:

> *a* helps *b* relative to *b*'s autonomously acquired goal to achieve $\gamma$ if and only if a) *a* intends to contribute to *b*'s achieving $\gamma$ and carries out this intention, and b) *b* accepts a). (Tuomela 2000, p. 136; adapted notation).

It is important to point out at once that Tuomela himself sees his characterization as too strong (Tuomela 2000, p. 136). We are seeking specifically a more basic, and, at the same time, more general notion. We would like to contrast it with the two conditions provided by Tuomela.

First, let us consider the point of view of the helper *a*, that is the a) condition of Tuomela's definition. As we have already said, the emphasis in our framework is not on the intentional notion of goal of an agent, it is rather on the more general notion of end (*telos*) of an action. With this in mind, not to be committed to strictly purposeful actions can also leave room for helping behavior in other forms of agency. For example, what appears to be a spontaneous tendency of children to cooperate (Warneken and Tomasello 2009) could be seen as an impulsive helping behavior:

> The behavior is as simple as it is surprising—and it is highly robust. Drop an object accidentally on the floor and try to reach for it, for example, from a desk, and infants as young as 14–18 months of age will toddle over, pick it up and return it to you. (Warneken and Tomasello 2009, p. 397).

This obviously depends on what position one may take with respect to intentions. If impulses are considered as intentions, then Tuomela's definition is valid, in this respect at least. If, instead, we are not willing to accept impulses as some form of intentions, then our teleological notion of help is more flexible, since it covers both options. But there are other cases that do not fit Tuomela's definition in any way. Consider, for example, a competitive game, where some unintentional behavior of some player just helped the opponent in taking advantage in the game. For instance, a football kicked by an attacker and bouncing off a defender into the net. Out of all the possible positions on the field, the defender chose this one. It is a controlling agency of the defender that is ill-fatedly directed towards making sure that the opponent's attempt to score is realized. Finally, one can imagine cases of help also in actions where it is difficult to assess if they are intentional or unintentional, as in side effects or lucky actions (Mele 2003). These cases exclude to us as requirements both the adoption of someone's goal and the formation of an intention. In such cases the teleological stance that we adopt shows instead its benefits.

Secondly, the other point of view to consider is the helpee's one. The first condition regarding *b* is that *b*'s goal has to be "autonomously acquired". This assumption is meant, as Tuomela himself states, to exclude cases where *a* coerces *b* to have the goal $\gamma$ and *b* accepts the "help" in virtue of that coercion. This condition seems significant also in the light of what we said about weak interpersonal control, that also applies to our conception of help. As we stated in Sect. 2, the interpersonal control we are interested in is neither a form of mind control nor some way of bringing about that the helpee brings about that something is the case. This amounts to say that the control provided in a helping behavior has to be over a conditional state of affairs, whose antecedent is a proper volition/trying of the helpee. (Classically, this conditional is also a disjunctive state of affairs where one disjunct is the negation of the trying.) The relevance of the helpee's rational volition is also the primary assumption taken by Chisholm and Zimmerman (in an otherwise mysterious working note):

> My being helped by someone to bring about some event implies an intentional relation between me and the event in question. Jones's helping Robinson to do something implies that Robinson, at least, "knows what he's doing", whether or not Jones does. (Chisholm and Zimmerman 1996, p. 402)

The other condition imposed by Tuomela that regards *b*, the acceptance condition—that is, the requirement that *b* accepts that *a* intends to contribute to *b*'s achieving $\gamma$ and carries out this intention—is instead more problematic. There are many cases of help where the acceptance is not needed, because the controlling agency fits, so to say, with the volition of the helpee, with no need for the helpee to have any agreement on it. Consider cases of paternalistic help, that is a rather common manifestation of help in human behavior. Recent studies in developmental psychology show how when facing the situation where an experimenter requests something that is ill-suited to achieving their ultimate goal, 3-year-old children override the request in favor of what they believe is best for them (Martin and Olson 2013). There is no acceptance from the helpee and yet, help occurs. The same goes in our previous football example where the defender unwittingly helps the attacker to score. The attacker is not expected, in order to be successfully helped, to accept what the unlucky defender's agency is going to provide him. Even if the helpee is unaware of the helper's agency, it is sufficient for him to take advantage of the situation, and the resulting event can be considered help.

Given these observations we can now focus on our notion of *successful help*:

> *a helps b relative to b's trying to bring about that $\phi$, if and only if: (i) a brings it about that: if b tries to bring about that $\phi$ then $\phi$ is the case; (ii) b tries to bring about that $\phi$.*

Condition (i) represents the controlling agency of help. It is a bringing about, so it is an *effective* guidance interpersonal control. Its content is a material implication, dependent on the helpee's volition. As the left hand side of the condition is exactly $b$'s attempt to bring about $\phi$, we will qualify it as a *justified assistance*. We can have a successful case of help when this controlling agency properly combines with condition (ii), the controlled agency. It requires that the helpee has to actually try to bring about that $\phi$. We will qualify this property as an *opportune assistance*. As these two agencies are properly aligned, we are in presence of a *well-situated* interpersonal control, formally defined in the previous section. It implies that help is *successful* and $\phi$ is the case.

Since the controlling agency is conditional, agent $a$ can help $b$ without necessarily actively intervening in the situation. Agent $a$'s agency may be decisive for the truth of $\phi$ or may be redundant. We will define in particular the fact that the assistance is *decisive* when $b$ does not bring about $\phi$ himself.

Elgesem's example is exactly about this. He assists his son for the sake of his son's attempt to be fed. If the baby is able to do it by itself the assistance is not decisive, otherwise, Elgesem makes sure that the baby gets fed when it tries. The example of the keys mentioned in the introduction, is also in line with our definition. If we see our partner about to open door, as we reach for the keys in our pocket ready to open the door, we help her anyway, even if she does get the keys first and does open the door. Help is exactly about the preparedness to provide to the helpee, what is needed in order to accomplish what she is trying to accomplish. It is also what Elgesem calls a back-up system in his example. The preparedness is a kind of guidance interpersonal control. But, we want to emphasize it, this control is *weak*. First, it is not mind control or some other strong way to bring about that the helpee brings about that $\phi$ holds. Agent $b$ can autonomously acquire her volition $A_b\phi$. Second, it is a form of guidance. The state of affairs $\phi$ could become true even without an active participation of the helper. And even if his intervention is decisive, $\phi$ has to become true only with the precondition of the helpee's volition.

# 5 Logical Aspects of Help

## *5.1 Some Formal Properties of Assisting Behavior*

On our way to characterize the notion of successful help that we defended before, we propose a few properties pertaining to what we may call more generally *assisting behaviors*, or simply *assistances*.

To start with, paternalism is a limiting factor to the meaningfulness of help. It occurs when the controlling agency of $a$ does not properly capitalize on an attempt of $b$ to bring about $\gamma$. We then start by characterizing a condition for the controlling agency to be a justified assistance (of $a$ towards $b$).

**Definition 4 (JA).** For any $\psi$, a guidance interpersonal control $\mathsf{GIC}(X_1^{ab}, X_2^{ab}, X_3^{ab}, \psi)$ is a *justified assistance* when

$$\vdash X_2^{ab}\psi \to A_b\psi$$

We say that an interpersonal control is a justified assistance when the mode of the content of the controlling agency at least includes an attempt of *b*.

All assistances are not necessary for bringing about the state of affairs sought after by a controlled agency. One class of these assistances is that of faked assistances. They occur when the mode of the content of the controlling agency at least includes *b* bringing about its volition.

**Definition 5 (FA).** For any $\psi$, a guidance interpersonal control $\mathsf{GIC}(X_1^{ab}, X_2^{ab}, X_3^{ab}, \psi)$ is a *faked assistance* when

$$\vdash X_2^{ab}\psi \to E_b\psi$$

So the controlling agency of a faked assistance is over a state of affairs satisfying $\gamma$ when, at least, *b* does bring about $\gamma$.

Critical properties of interpersonal control depend on the controlled agency. A decisive assistance occurs when the controlled agency of an interpersonal control for $\gamma$ implies that *b does not* already bring about $\gamma$.

**Definition 6 (DA).** For any $\psi$, a guidance interpersonal control $\mathsf{GIC}(X_1^{ab}, X_2^{ab}, X_3^{ab}, \psi)$ is a *decisive assistance* when

$$\vdash X_3^{ab}\psi \to \neg E_b\psi$$

This is decisive in the sense that *b* does not bring about $\gamma$ himself. This is not necessarily decisive in the sense that $\gamma$ would not be true if it were not for *b*'s action. Indeed, $\gamma$ might be true coincidentally for some reason independent of *a* and *b*'s actions. One can of course define a stronger property as follows:

**Definition 7 (SD).** For any $\psi$, a guidance interpersonal control $\mathsf{GIC}(X_1^{ab}, X_2^{ab}, X_3^{ab}, \psi)$ is a *strongly decisive assistance* when

$$\vdash X_3^{ab}\psi \to \neg\psi$$

Surely however, assistances would barely deserve the name if it were not for *b* to actually try to bring about a state of affairs.

**Definition 8 (OA).** For any $\psi$, a guidance interpersonal control $\mathsf{GIC}(X_1^{ab}, X_2^{ab}, X_3^{ab}, \psi)$ is an *opportune assistance* when

$$\vdash X_3^{ab}\psi \to A_b\psi$$

In an opportune assistance for $\gamma$, the controlled agency at least implies that $b$ tries to bring about $\gamma$.

## 5.2 A Simple Account of Successful Help

Finally, successful help (of $a$ towards $b$ for $\gamma$) can be rigorously defined as the weakest form of well-situated opportune effective assistance. That is, $\mathsf{GIC}(X_1^{ab}, X_2^{ab}, X_3^{ab}, \gamma)$, where:

1. $\vdash X_1^{ab}\psi \leftrightarrow E_a\psi$
2. $\vdash X_2^{ab}\psi \leftrightarrow A_b\psi$
3. $\vdash X_3^{ab}\psi \leftrightarrow A_b\psi$

It is worth defining a new dedicated modality. Thus, we obtain:

$$[a:b]\gamma \overset{\text{def}}{=} E_a(A_b\gamma \to \gamma) \wedge A_b\gamma$$

which we read "$a$ successfully helps $b$ to bring about $\gamma$".

It is *successful* because we have the following expected property by applying (success) and (prop):

**Proposition 2.** $\vdash [a:b]\phi \to \phi$

It is an assistance for three reasons. First, there is an *assistee*. It is a volition of $b$ to bring about $\gamma$ and $b$ does try. Second, there is a *assistant*. the content of $a$'s control is over the state of affairs where $\gamma$ is true whenever $b$ tries to bring about $\gamma$. Hence, $i$'s guidance is reactive to $b$'s goodwill in the action. Third, it is compelling to a formalization of assistance that $[a:b]\gamma \wedge \neg E_a\gamma \wedge \neg E_b\gamma$ is a consistent formula. That is, it is possible that $a$ successfully helps $b$ to bring about $\gamma$, and still, neither $a$ nor $b$ brings about $\gamma$. Hence, the success of the assistance described by $[a:b]\gamma$ comes from some cohesion between $a$ and $b$.

**Elgesem's example.** Back to Elgesem's example about his one-year old boy (see Sect. 2). There are two cases: "sometimes [the boy] succeeds in getting the food into his mouth with the spoon, and sometimes not." When he does succeed, Elgesem argues that the boy does bring about that he has food in his mouth. That is, $E_{boy}\mathsf{food}$, where $\mathsf{food}$ stands for "the boy has food in his mouth". When he does not succeed however, there is a "back-up system". It is, we argued, the help provided by the father. Note that the accent is put on the boy being in the process of learning to eat by himself. There is no case of feeding the boy against his will. So, we must say that indeed the boy tries to bring about that he has food in his mouth: $A_{boy}\mathsf{food}$. It is the controlled agency. The back-up system is the controlling agency, which consists in making sure that the boy has food in his mouth when he tries to bring about that the food is in his mouth. The agent of the controlling agency is Dag Elgesem himself, so we have: $E_{dag}(A_{boy}\mathsf{food} \to \mathsf{food})$. This is a case of effective, well-situated opportune guidance interpersonal control of Dag over the boy's agency towards the boy having food in his mouth.

To sum up, at least one of the following holds:

- $E_{boy}\mathsf{food}$
- $E_{dag}(A_{boy}\mathsf{food} \to \mathsf{food}) \wedge A_{boy}\mathsf{food}$

which implies that $\mathsf{food}$ holds no matter what.

### 5.3 Proven Properties of Interpersonal Control and Assistances

The logical theory allows to reason about more complex properties of contextual agency now expressible with our vocabulary. Some properties are expected from the choice of terminology. We can verify for instance that a strongly decisive assistance is a decisive assistance.

**Theorem 1.** *Strongly decisive assistance is decisive.*

*Proof.*

1. $\{SD\} \vdash X_3^{ab}\gamma \to \neg\gamma$ (from SD)
2. $\{SD\} \vdash \neg\gamma \to \neg E_b\gamma$ (from (success) and (prop))
3. $\{SD\} \vdash X_3^{ab}\gamma \to \neg E_b\gamma$ (from 1., 2., and (prop))
4. $\{SD\} \vdash DA$ (from 3. and DA)

But typically, properties are not so transparent. We prove two more theorems.

**Theorem 2.** *Opportune well-situated assistance is justified.*

*Proof.*

1. $\{OA, WS\} \vdash (X_3^{ab}\gamma \to A_b\gamma) \wedge (X_3^{ab}\gamma \leftrightarrow X_2^{ab}\gamma)$ (from OA, WS, and (prop))
2. $\{OA, WS\} \vdash X_2^{ab}\gamma \to A_b\gamma$ (from 1. and (prop))
3. $\{OA, WS\} \vdash JA$ (from 2. and JA)

**Theorem 3.** *Effective faked assistance is impossible.*

*Proof.*

1. $\{EF, FA\} \vdash X_1^{ab}\gamma \leftrightarrow E_a\gamma$ (from EF)
2. $\{EF, FA\} \vdash X_2^{ab}\gamma \to E_b\gamma$ (from FA)
3. $\{EF, FA\} \vdash X_2^{ab}\gamma \to \gamma$ (from 2., (success), and (prop))
4. $\{EF, FA\} \vdash E_a(X_2^{ab}\gamma \to \gamma) \to \bot$ (from 3., (notaut), and (prop))
5. $\{EF, FA\} \vdash X_1^{ab}(X_2^{ab}\gamma \to \gamma) \to \bot$ (from 1., 4., and (prop))
6. $\{EF, FA\} \vdash X_1^{ab}(X_2^{ab}\gamma \to \gamma) \wedge X_3^{ab}\gamma \to \bot$ (from 5. and (prop))
7. $\{EF, FA\} \vdash \mathsf{GIC}(X_1^{ab}, X_2^{ab}, X_3^{ab}, \gamma) \to \bot$ (from 6. by definition)

In English: effective faked assistance is impossible because it occurs when (i) the agentive mode of the assistant is to actually bring about the content of the interpersonal control (for $\gamma$), and (ii) the controlled agency includes the fact that the assistee already brings about $\gamma$. But by (ii) and (success), the content of the controlling agency is trivial: it is a theorem in the logic. But by axiom (notaut), the logic does not allow an agent to bring about tautologies, which is what the assistant's mode is by (i).

On the other hand, tentative faked assistance is possible. The reason is rather ordinary: according to our axiomatics of BIAT, it is possible for an agent to attempt to bring about tautologies.

## 6 Subjective Help

We have been arguing for and formalizing an account of help which is unilateral and triggered by an attempt. It is unilateral because we can provide our help to someone without her accepting it. She could be unaware of our actions, or she could be unwilling to receive it. Help is based on trying because it is agent $b$ (supposedly) trying to do something that triggers $a$'s action of help.

Here, we want to add that subjectivity plays a crucial role for characterizing an event as an event of help. Help is subjective since in helping $b$, agent $a$ can have imperfect information about $b$'s volition. There was a Norwegian TV commercial for *Japp* chocolate bars where a man finishes a jog on a mountain road and arrives panting at his sport car parked near a cliff. He proceeds to stretch, hands on the car, facing the cliff. With a background of Caribbean music, a rastaman is driving by, eating a chocolate bar. (The slogan says that it gives extra energy.) He sees the scene, and looking determined he stops his truck, jumps out, walks to the car and pushes it over the cliff. As the rastaman believed that the car owner was trying to push his car off the cliff, there is an aspect of helping behavior in this event.

We extend the BIAT framework with one modality $Bel_i$ for each agent $i$. The formula $Bel_i\phi$ reads that the agent $i$ believes $\phi$. Since our basis framework of agency is very abstract (BIAT is a weak modal logic), we do not assume much about the logic of $Bel_i$.

Any logic between S5 (full blown knowledge (Halpern and Moses 1992)), and the minimal modal logic should be consistent with our analysis in this paper. (Intermediate systems can be found in Hendricks and Symons (2014).) Although we will not do specific reasoning about beliefs in this paper, it is typically judicious for a work in modal logic to assume the following:

$$\text{(reb) if} \vdash \phi \leftrightarrow \psi \text{ then} \vdash Bel_i\phi \leftrightarrow Bel_i\psi$$

We need a feasible methodology to pick out *events of subjective help* out of the many types of weak interpersonal control. We must concede that we cannot think of a unique methodology that would explain satisfyingly and completely why we consider that some instance of interpersonal control is not an event of help and why we consider some other instance as a typical event of help. Nonetheless, we can reiterate what aspects we see as relevant, propose the pertinent sets of parameters (viz., $X_2^{ab}$ and $X_3^{ab}$) and exhaustively analyze their possible combinations.

The relevant aspects of subjective help are:

- it is based on a (presumed) attempt on the assistee part;
- it is subjective on the assistant part.

Assuming that the relevant beliefs of the assistant concern the trying events of the assistee, this considerably restricts our research space. Finally, we only consider help as an effective agency. Thus we adopt EF, meaning that $X_1^{ab}$ has to be $E_a$. We will later comment on replacing effective agency by attempt agency.

**Identifying the relevant subjective events of effective help.** With the previous considerations in mind, we will look specifically at the cases of interpersonal control $\mathsf{GIC}(X_1^{ab}, X_2^{ab}, X_3^{ab}, \gamma)$ where the mode of the controlled agency $X_3^{ab}$ and the mode of the content of the controlling agency $X_2^{ab}$ can obey one of three possible principles. We will examine the following (for all nine combinations of $X = X_2^{ab}$ and $X = X_3^{ab}$):

- $\vdash X\psi \leftrightarrow A_b\psi$
- $\vdash X\psi \leftrightarrow Bel_a A_b\psi$
- $\vdash X\psi \leftrightarrow A_b\psi \wedge Bel_a A_b\psi$

For clarity of exposition we will use several variations on a toy scenario of interaction between agent $a$ and agent $b$, where $a$ operates two push-buttons 1 and 2, and $b$ operates a push-button 3. A light is on (property captured by $\gamma$) iff 1 is pressed, and at least one of 2 and 3 is pressed. Suppose that only agent $b$ may have some concern over $\gamma$, and pushes his button 3 as a way to try to bring about that the light is on: $A_b\gamma$. Agent $a$ can assist $b$ in doing so, but may have imperfect knowledge as to whether $b$ indeed tries to bring about $\gamma$. Either $a$ believes that $b$ tries to bring about $\gamma$ ($Bel_a A_b\gamma$) or does not ($\neg Bel_a A_b\gamma$).

When the mode of the content of the controlling agency is $A_b$, the interpersonal control is (minimally) justified. In the lights of our toy scenario, the controlling agency may be seen as $a$ indiscernibly pressing button 1, no matter what his beliefs are. If $b$ tries to bring about $\gamma$, thus pressing the button 3, $\gamma$ would hold.

- $E_a(A_b\gamma \rightarrow \gamma) \wedge A_b\gamma$. It is precisely our account of successful help: effective, opportune and well-situated interpersonal control.
- $E_a(A_b\gamma \rightarrow \gamma) \wedge Bel_a A_b\gamma$. It is not (necessarily) an opportune assistance. It also does not ensure that $\gamma$ indeed holds. Agent $a$ believes that $b$ tries to bring about $\gamma$, but this belief is not taken into account in the controlling agency.
- $E_a(A_b\gamma \rightarrow \gamma) \wedge (A_b\gamma \wedge Bel_a A_b\gamma)$. It is logically equivalent to the conjunction of the two previous cases. It is an effective, opportune and well-situated interpersonal control, and agent $a$'s belief does not add anything remarkable.

When the mode of the content of the controlling agency is $Bel_a A_b$ we face a subjectively sensitive case of controlling agency. It is not justified. In our scenario, the controlling agency may be seen as $a$ pressing the button 1 no matter what, and also pressing 2 whenever he believes that $b$ tries to bring about $\gamma$.

- $E_a(Bel_aA_b\gamma \rightarrow \gamma) \wedge A_b\gamma$. Although it is an opportune assistance, it does not (necessarily) imply that $\gamma$ holds.
- $E_a(Bel_aA_b\gamma \rightarrow \gamma) \wedge Bel_aA_b\gamma$. It is not (necessarily) an opportune assistance but $\gamma$ holds.
- $E_a(Bel_aA_b\gamma \rightarrow \gamma) \wedge (A_b\gamma \wedge Bel_aA_b\gamma)$. It is logically equivalent to the conjunction of the two previous cases. It is an opportune assistance, and the agent $a$'s belief has the effect that the interpersonal control results in $\gamma$ being true.

*Remark 2.* We can observe that our description of the variants of the toy example suggests that in the previous second and third cases $a$ presses both push-buttons 1 and 2. For all practical purpose we might say, in this example, that $a$ does bring about that $\gamma$. We prefer to leave the question open in this paper whether it should be a general principle, . Possibly, it could be argued that $(E_a(Bel_aA_b\gamma \rightarrow \gamma) \wedge Bel_aA_b\gamma) \rightarrow E_a\gamma$ would make a pertinent principle of agency.

When the mode of the content of the controlling agency is $A_b\psi \wedge Bel_aA_b\psi$, the interpersonal control is justified, and the controlling agency is subjectively sensitive. In the scenario, the controlling agency may then be seen as the variant where $a$ presses the push-button 1 whenever he believes that $b$ tries to bring about $\gamma$.

- $E_a((A_b\gamma \wedge Bel_aA_b\gamma) \rightarrow \gamma) \wedge A_b\gamma$. It is an opportune assistance, but it is not (necessarily) true that $\gamma$.
- $E_a((A_b\gamma \wedge Bel_aA_b\gamma) \rightarrow \gamma) \wedge Bel_aA_b\gamma$. It is not (necessarily) an opportune assistance, and it is not (necessarily) true that $\gamma$.
- $E_a((A_b\gamma \wedge Bel_aA_b\gamma) \rightarrow \gamma) \wedge (A_b\gamma \wedge Bel_aA_b\gamma)$. It is logically equivalent to the conjunction of the two previous cases. It is an opportune assistance, $a$ justifiably believes that $b$ tries to bring about $\gamma$. It does imply that $\gamma$ holds.

**Tentative subjective help.** Each case of effective help that we just mentioned naturally has a counterpart as tentative help.

In order to talk conveniently about tentative subjective help, we must come up with an adequate modification of the toy scenario used previously. Agent $a$ now is at some distance from the push-buttons 1 and 2, and has to throw skillfully a juggling ball at each of them in order to activate them. What is important here is that unlike pushing a button, the result of throwing a ball at a button has a non-deterministic result. Throwing a juggling ball at a button, we consider it as a trying to press the button. Agent $b$ still operates the push-button 3, normally as before. In addition, the light is on in the same conditions as before, that is, when 1 is pressed, and at least one of 2 and 3 is pressed.

Consider again the three cases:

- $\vdash X_2^{ab}\psi \leftrightarrow A_b\psi$
- $\vdash X_2^{ab}\psi \leftrightarrow Bel_aA_b\psi$
- $\vdash X_2^{ab}\psi \leftrightarrow A_b\psi \wedge Bel_aA_b\psi$

but under the assumption TE this time. The controlling agency of our modified toy scenario can then be described respectively as:

- $A_a(A_b\gamma \rightarrow \gamma)$: $a$ indiscernibly throws a juggling ball at button 1, no matter what his beliefs are.
- $A_a(Bel_a A_b\gamma \rightarrow \gamma)$: $a$ throws a juggling ball at button 1 no matter what, and also throws a juggling ball at button 2 whenever he believes that $b$ tries to bring about $\gamma$.
- $A_a((A_b\gamma \wedge Bel_a A_b\gamma) \rightarrow \gamma)$: $a$ throws a juggling ball at the push-button 1 whenever he believes that $b$ tries to bring about $\gamma$.

Finally, essentially the same comments would be made about the resulting interpersonal controls, except that none of them would (necessarily) imply that $\gamma$ holds.

**Warneken and Tomasello's experiments.** In Warneken and Tomasello (2006), Warneken and Tomasello describe four experiments of help behavior in prelinguistic or just-linguistic children. In one of them, the adult tries, or at least act as he tries, to put magazines into a cabinet. But the doors are closed and he bumps into it instead. The experiment[8] shows that the infant helps the adult to achieve his task by opening the doors.

Say that $\mathsf{open}$ stands for the "cabinet is open". The scenario can be formalized in the logic.

1. The subjective controlled agency: $Bel_{toddler}A_{adult}\mathsf{open}$
2. The subjective controlling agency: $E_{toddler}(Bel_{toddler}A_{adult}\mathsf{open} \rightarrow \mathsf{open})$
3. Possibly: $A_{adult}\mathsf{open}$

So, (1) the toddler believes that the adult tries to bring about that the cabinet is open, and (2) the toddler brings about that the cabinet is open when he believes that the adult tries to bring about that the cabinet is open. The subjective help captured by the experiment is then an interpersonal control

$$\mathsf{GIC}(E_{toddler}, Bel_{toddler}A_{adult}, Bel_{toddler}A_{adult}, \mathsf{open}).$$

It is a well-situated and effective interpersonal control. Also, it is a successful subjective help in the sense that $\vdash \mathsf{GIC}(E_{toddler}, Bel_{toddler}A_{adult}, Bel_{toddler}A_{adult}, \mathsf{open}) \rightarrow \mathsf{open}$. (3) It is irrelevant whether the adult indeed tries to bring about that the cabinet is open, and the setting of the experiment does not allow us to conclude any way or the other. Hence, it is not an opportune event of assistance.

**Acknowledgements** Emanuele Bottazzi was supported by the "Postdoc 2011" funding programme of the Autonomous Province of Trento, Italy (project "STACCO"). Nicolas Troquard was supported by a Marie Curie fellowship under grant PCOFUND-GA-2008-226070 (project "LASTS").

[8] Captured in video http://www.eva.mpg.de/psycho/videos/children_cabinet.mpg

## References

Belnap, Nuel, and M. Perloff. 1988. Seeing to it that: A canonical form for agentives. *Theoria* 54(3): 175–199.

Belnap, Nuel, Michael Perloff, and Ming Xu. 2001. *Facing the future (agents and choices in our indeterminist world)*. Oxford/New York: Oxford University Press.

Bratman, Michael E. 1992. Shared cooperative activity. *The Philosophical Review* 101(2): 327–341.

Chisholm, Roderick M., and Dean W. Zimmerman. 1996. On the logic of intentional help. *Faith and Philsophy* 13(3): 402–404.

Darley, John. M., and Bibb Latané. 1968. Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology* 8(4): 377–383.

Elgesem, Dag. 1993. *Action theory and modal logic*. PhD thesis, Universitetet i Oslo.

Elgesem, Dag. 1997. The modal logic of agency. *Nordic Journal of Philosophical Logic* 2(2): 1–46.

Fischer, John Martin. 1994. *The metaphysics of free will*. Oxford: Blackwell.

Fischer, John Martin. 2012. *Deep control*. Oxford/New York: Oxford University Press.

Frankfurt, Harry. 1969. Alternate possibilities and moral responsibility. *Journal of Philosophy* 66: 829–839.

Frankfurt, Harry. 1988. *The importance of what we care about*. Cambridge/New York: Cambridge University Press.

Gelati, Jonathan, Antonino Rotolo, Giovanni Sartor, and Guido Governatori. 2004. Normative autonomy and normative co-ordination: Declarative power, representation, and mandate. *Artificial Intelligence and Law* 12: 53–81.

Governatori, Guido, and Antonino Rotolo. 2005. On the axiomatisation of Elgesem's logic of agency and ability. *Journal of Philosophical Logic* 34(4): 403–431.

Halpern, Joseph Y., and Yoram Moses. 1992. A guide to completeness and complexity for modal logics of knowledge and belief. *Artificial Intelligence* 54(2): 319–379.

Hendricks, Vincent, and John Symons. 2014. Epistemic logic. In *The Stanford encyclopedia of philosophy*, ed. Edward N. Zalta, Spring 2014 edition. CSLI, Stanford University. http://plato.stanford.edu/cgi-bin/encyclopedia/archinfo.cgi?entry=freewill

Hilpinen, Risto. 1997. On action and agency. In *Logic, action and cognition: Essays in philosophical logic*, ed. E. Ejerhed and S. Lindström, 3–27. Dordrecht: Kluwer.

Hilpinen, Risto. 2001. Stig Kanger on deontic logic. In *Collected papers of Stig Kanger with essays on his life and work*, vol. 2, ed. Ghita Holmström-Hintikka, Sten Lindstrom, and Rysiek Sliwinski. Dordrecht/Boston: Kluwer.

Hornsby, Jennifer. 1980. *Actions*. London: Routledge and Keegan.

Hornsby, Jennifer. 2010. Trying to act. In *A companion to the philosophy of action*, ed. Timothy O'Connor and Constantine Sandis, 18–25. Chichester/Malden: Wiley-Blackwell.

Horty, John F. 2001. *Agency and deontic logic*. Oxford: Oxford University Press.

Kanger, Stig. 1957/1971. New foudations for ethical theory. In *Deontic logic: Introductory and systematic readings*, ed. Risto Hilpinen, 36–58. Dordrecht: Reidel.

Kanger, Stig. 1972. Law and logic. *Theoria* 38: 105–132.

Kanger, Stig, and Helle Kanger. 1966. Rights and parliamentarism. *Theoria* 32: 85–115.

Kilpatrick, Shelley Dean. 2007. Helping behaviour. In *Encyclopedia of social psychology*, vol. 2, ed. Roy F. Baumeister and Kathleen D. Vohs, 420–424. Thousand Oaks: SAGE Publications.

Lindahl, Lars. 1977. *Position and change – A study in law and logic*. Dordrecht/Boston: D. Reidel.

Lorini, Emiliano, and Andreas Herzig. 2008. A logic of intention and attempt. *Synthese* 163: 45–77.

Martin, Alia, and Kristina R. Olson. 2013. When kids know better: Paternalistic helping in 3-year-old children. *Developmental Psychology* 49(11): 2071–2081.

Mele, Alfred R. 2003. Intentional action: Controversies, data, and core hypotheses. *Philosophical Psychology* 16: 325–340.

Meyer, Susan Sauve. 2006. Aristotle on the voluntary. In *The Blackwell guide to Aristotle's Nicomachean ethics*, ed. Richard Kraut, 137–157. Malden/Oxford: Blackwell.

Miller, Seumas. 2002. *Social action: A teleological account*. Cambridge: Cambridge University Press.

Milligram, Elijah. 2010. Pluralism about action. In *A companion to the philosophy of action*, ed. Timothy O'Connor and Constantine Sandis, 90–96. Chichester/Malden: Blackwell.

O'Connor, Timothy. 2014. Free will. In *The Stanford encyclopedia of philosophy*, ed. Edward N. Zalta, Fall 2014 edition.

O'Shaughnessy, Brian. 1980. *The will*. Cambridge: Cambridge University Press.

Pearce, Philip L., and Paul R. Amato. 1980. A taxonomy of helping: A multidimensional scaling analysis. *Social Psychology Quarterly* 43(4): 363–371.

Pörn, Ingmar. 1977. *Action theory and social science: Some formal models*, Synthese library 120. Dordrecht: D. Reidel.

Royakkers, Lambèr. 2000. Combining deontic and action logics for collective agency. In *Legal knowledge and information systems. Jurix 2000: The thirteenth annual conference*, ed. Joost Breuker, Ronald Leenes, and Radboud Winkels, 135–146. Amsterdam/Washington, DC: IOS Press.

Santos, Felipe, and José Carmo. 1996. Indirect action, influence and responsibility. In *Proceedings of DEON'96*, 194–215. London: Springer.

Santos, Felipe, Andrew Jones, and José Carmo. 1997. Responsibility for action in organisations: A formal model. In *Contemporary action theory*, vol. 1, ed. G. Holmström-Hintikka and R. Tuomela, 333–348. Dordrecht/Boston: Kluwer.

Searle, John. R. 1990. Collective intentions and actions. In *Intentions in communication*, ed. P. Cohen, J. Morgan, and M. E. Pollack. Cambridge: MIT Press.

Smithson, Michael, and Paul Amato. 1982. An unstudied region of helping: An extension of the Pearce-Amato cognitive taxonomy. *Social Psychology Quarterly* 45(2): 67–76.

Sommerhoff, Gerd. 1969. The abstract characteristics of living systems. In *Systems thinking: Selected readings*, ed. F.E. Emery. Harmonsworth: Penguin.

Troquard, Nicolas. 2014. Reasoning about coalitional agency and ability in the logics of "bringing-it-about". *Autonomous Agents and Multi-agent Systems* 28(3): 381–407.

Tuomela, Raimo. 2000. *Cooperation – A philosophical study*. Dordrecht: Springer.

Tuomela, Raimo. 2007. *The philosophy of sociality: The shared point of view*. Oxford/New York: Oxford University Press.

Warneken, Felix, and Michael Tomasello. 2006. Altruistic helping in human infants and young chimpanzees. *Science* 311(5765): 1301–1303.

Warneken, Felix, and Michael Tomasello. 2009. The roots of human altruism. *British Journal of Psychology* 100: 445–471.

# Healing Social Sciences' Psycho-phobia: Founding Social Action and Structure on Mental Representations

**Cristiano Castelfranchi**

**Abstract** I first argue against the "psycho-phobia" that has characterized the foundation of the social sciences and invalidates many social policies. I then present a basic ontology of social actions by examining their most important forms, with a special focus on pro-social actions, in particular Goal Delegation and Goal Adoption. These action types are the basic atoms of exchange, cooperation, group action, and organization. The proposed ontology is grounded in the mental representations (beliefs and goals) of the agents involved in social (inter)actions: the individual social mind. I will argue that such an analytical account of social action is needed to provide an adequate conceptual apparatus for social theory. In particular, I will try to show why we need to consider mind-reading and cognitive agents (and therefore, why we have to study the cognitive underpinnings of coordination and social action); why we need to consider agents' goals about the mind of others in interaction and collaboration, as well to explain group loyality and social commitment to the other; why cognition, communication and agreement are not enough for modeling and implementing cooperation; why emergent pre-cognitive structures and constraints should be formalized; why emergent cooperation is also needed among planning and deliberative social actors; and why also the Nets with their topological structure and dynamics are in fact mind-based.

**Keywords** Psyco-phobia • Social action • Social mind • Mind-reading • Cooperation • Emergent cooperation • Functions

C. Castelfranchi (✉)
Institute of Cognitive Sciences and Technologies – CNR, Rome, Italy
e-mail: cristiano.castelfranchi@istc.cnr.it

A. Herzig, E. Lorini (eds.), *The Cognitive Foundations of Group Attitudes and Social Interaction*, Studies in the Philosophy of Sociality 5,
DOI 10.1007/978-3-319-21732-1_2

# 1 Introduction: "Psycho-phobia" and the "Cognitive Mediators" of Social Phenomena

I would like to begin with a few remarks on the "psycho-phobia" that, I believe, has characterized much of the social sciences (SSs)[1] since their beginnings, and that invalidates many social policies derived from social science theories. Let me start with the great von Hayek.

Hayek, to be sure, was very much interested in psychology and in the psychological foundation of action and knowledge, as well as in the link between economics and psychology. He was thus not himself psychophobic. However, it is also true (in my view[2]) that Hayek is the scholar who has most directly and clearly expressed a very wide-spread view about the goals of the SSs, that has even become a "foundational" view for many, according to which the SSs should study the complex effects of human collective behavior that are not mentally represented – those which are neither intended nor understood. In Hayek's view, if all the aspects and consequences of human actions where understood or intended, then the SSs would not exist, because there would be no need for them – psychology would be enough.

Just a citation:

> This problem (the spontaneous emergence of an unintentional social order and institutions) is in no way specific of the economic science . . . it doubtless is *the core theoretical problem of the whole social science* (von Hayek, *Knowledge, Market, Planning*).

In such a way, Hayek identifies (in my view correctly) the focal point, the central issue and the *raison d'être* of the SSs as consisting in the challenge of *going beyond minds*, their understanding and control. Hayek is deeply interested in the connection between economic theory and psychological theory, but he does not regard our minds as the foundation and the reason for the SSs; rather, he sees *their privileged object in those social phenomena that elude human intelligence and intention*.

To be sure, this view of the object of the SSs (social phenomena which go "beyond psychology") is based on an unilateral and unidirectional view of the connection between the mind and the extra-mental world; actually, what goes on beyond our minds also affects our minds – the link between the mind and the extra-

[1] Although certainly not all: Excluded are pre- and post-behaviorist social psychology, as well as some parts of sociology, economics and political science, particularly those concerned with phenomena such as marketing, propaganda and political demagogy. I am deliberately simplifying matters to bring into sharp relief a problem that is frequently not clearly perceived and whose importance is widely underestimated. Notice that here I will use a restricted notion of "social sciences", excluding on purpose psychology (at least the 'general' and 'cognitive' one); the sciences studing the sociological, collective structures, institutions, and behaviors.

[2] Note that I am not interested here in the history of ideas; I am only interested in the ideas – in capturing, using and discussing them. Hence, if Hayek didn't exactly say or mean what I attribute to him, and just is the current "Vulgate", this makes no difference to the present argument. What matters to me is the sin, not the sinner.

mental world is dialectic [bottom-up and top-down]. Because of this, the study of the sociality that exists in minds, and through minds, is not just a part of psychology (where it is studied under headings such as social psychology, mass psychology, socialization, social attitudes and emotions, social influence, etc.). We need not only a psychology of social action and social relations; we also need a *sociology of minds, an anthropology of minds, and an economics of minds*.

"Emergence" is in fact a dialectic, bidirectional dynamics: The emergent structure feeds back into minds and shapes and constructs them, and it does so *in order to* reproduce and stabilize itself. We therefore need not only a cognitive micro-level theory to ground the phenomena of the social sciences; we also need a macro-level causal theory of mental representations shaping and control on our behavior. Hence, social theory and social science are doubly bound to psychology, or better, to cognitive science: Social phenomena (both micro and macro) are both dependent on the minds of actors and construct the minds of actors, including those of the scientists.

The described *psycho-phobia* – the attempt, which is motivated in part by methodological (behaviorist) scruples and in part by the perceived need to establish the identity of the SSs – to ground the behavioral sciences in phenomena outside of the mind that controls them, and even to ignore the mind and its contents, is a widespread phenomenon, and is often repeated in the foundatational texts of the SSs. Let me illustrate this claim with a few examples.

This *psycho-phobia*, this identitarian and methodological need of founding the behavioral sciences outside the mind that control it and even ignoring mind and its contents, is rather frequent and iterated in the foundation of those sciences. Let me just remind here:

- In Economics, Pareto's proposal for an explicitly non-psychological foundation of preferences and utility has been widely adopted, relieving economics from the burden of investigating the real motivations, decisional mechanisms and beliefs of people, with their biases and extra-economic motives (Bruni and Sugden 2007).[3]
- In sociology and anthropology, we find Garfinkel's influential but contradictory view. On the one hand, Garfinkel founds social interaction and coordination

[3] In economics, an explicit treatment of goals has been suppressed by replacing it by a single, implicit goal: utility maximization. Hence, for example, evaluating options or their consequences means appreciating their utilities. But how can the utilities of consequences (apart from the utility of money, which is therefore the ideal good of economists) be determined, if not by relating them to the person's realized and non-realized goals (desires, needs, projects) and their subjective importance (value)? Likewise, "options" are options only relatively to a given goal. When/if eventually Economics is obliged to come back to psychology, and to accept the need for a psychological foundation of preferences in *motives*, it identifies psychology (beyond "rational" decision and action that is already and well accounted for by economics) with "subjective experience", with sensations (with the psychology of the 700 and 800), and search for a simplistic foundation of preferences and motives: *pleasure*; or more sophisticately and obscurely: *happyness* (Bruni and Sugden 2007).

in a very clear way on trust and "perceived normality"; on the other hand he claims, in the very same text, that *"Meaningful events are entirely and exclusively events in a person's behavioral environment, … Hence there is no reason to look under the skull since nothing of interest is to be found there but brains"* (Garfinkel 1963, p. 190). Nonetheless, he continuous to use terms like "doubt", "uncertainty", "worries", "assumption", "expectation", "perceived", "well-known" in his analysis – clearly mentalistic notions.

– In Game Theory, we have the simplifying postulate of (i) perfect and mutual knowledge and perfect rationality of the players; (ii) irrelevance of the specific and concrete objectives of the players, but relevance just of the global value and the ordinal positioning of the alternative moves.

Both of these assumptions are false for real human actors. The first assumption – which concerns actor's beliefs and reasoning – has today been abandoned in Behavioral and Cognitive game theory; the second one, however, has still not been entirely abandoned (apart from some gestures in explanations of the "reasons", i.e. motives, for playing irrationally). However, many game theorists still are not convinced that this problem can be solved by referring to a systematic, foundational theory of motivations and goals; in fact, there remains much skepticism among game theorists towards psychology – psychology is regarded as the discipline of the intangible (mind, "representations"), and is frequently identified with the study of subjectivity, personality, and individual differences. How, game theorists ask, can on ground a science, its predictions and practical applications, on something so mutable and impalpable, private, subjective and unformalized?[4]

The strength of economy's drive for "autonomy" from psychology is comparable only with that of psychology's striving to distance itself from from its mother discipline, philosophy, and with it, from analytical, conceptual and theoretical work. At present, this rejection finds a smart complicity on the other side, philosophy: Let's delegate the conceptual work to philosophy as its proper job, possibly to philosophers interested in the theories and empirical results of psychology and operational models (like the so-called "experimental philosophy"). This division of labor may be academically enjoyable, but is in my view wrong-headed. It is the job of psychologists to work on their concepts, and to provide clear definitions and distinctions that ground useful discussions and allow the reasonable untangling of data.

I am not in favor of empiricist and descriptive psychological and sociological accounts of human behavior that lack theoretical depth and generalization power. Rather, I favor abstract, ideal-typical, formal models of the mind, which are useful to different high-level theories in the social sciences. However, these models can

[4] What an old-fashioned view of psychology this is! Outdated even before the cognitivist revolution! One can understand how this conception of psychology (stemming from the phenomenological and introspective tradition) invites one to accept Behaviorism (like several economists do)—at least behaviors are observable. And in case of a perceived inescapable need for "mental" foundations, it seems better to skip psychology completely and directly connect to the (pseudo)concreteness of brain: neuro-economics, neuro-ethics, neuro-politics, etc.

no longer be as simplistic and anti-cognitive as those traditionally proposed in game theory and in economics. Currently, the view is spreading in economics that the available alternatives are, on the one hand, an abstract, theoretical and formal model of mind (identified with the decision-theoretic model of rationality), and on the other hand an experimental economics based on empirical findings and specific observational "laws" of human behavior. This view is simply wrong. Other principled, predictive, theoretical and formal approaches to the human social mind – in particular those developed within cognitive science – are available, and in fact these models promise to be most useful for theoretical explanation and modeling in the social sciences. Logic and computational modeling of mental contents and processes provide us with the means for constructing much more complex, abstract top-down models of the mind, while agent-based computer simulation of social phenomena provides an experimental basis for their validation.

The identity-preserving, anti-psychological barrier erected by the social sciences, which forms the basis and warranty of the autonomy of the SSs by providing a well-defined, unique territory for the discipline, is wrong-headed for several reasons. Of course, it is true that the SSs need an *autonomous foundation*; they cannot be reduced to psychology (just like psychology cannot be reduced to neuroscience); each layer of increasing complexity and organization of reality needs new concepts to describe it, a new ontology, and new "laws" specific to that layer. I am therefore against reduction understood as eliminativism. But I am in favor of "reduction" understood as the bringing-back ("ri-conduzione") of a dialectical foundation between different organization layers.

However here it seems that on the one side there is the human "action", consciously guided, and necessarily based on "knowledge", rational, intentional, and effective (realizing the expected outcomes); on the *other* side, the not understood effects, not intended, not predictable and manageable ("spontaneous"); necessarily self-organizing and emergent, produced not by the mind or the minds but by the "invisible hand".

Indeed: even intentional human actions are not guided by knowledge and rationality but by beliefs, assumptions, opinions, illusions, ideologies, prejudices, cultural schemes and norms, values, cultural aims, including "impressions" about the complex emergent trends; and by heuristics and systematic distortions in reasoning and decision-making.

How much these epistemic and motivational representations that regulate our intentional conduct are *due to* the macro sociological, economic, anthropological, political levels? How the ones *are functional to the others*, not just mere complex effects and consequences?

That is, how can spontaneous order not just emerge from our autonomous acts, but maintain and reproduce itself without actively influencing and reproducing these acts and hence – because they are due to our cognitive representations and processes – by selecting and reproducing these cognitive mechanisms? The proposed answer is that *the Invisible Hand works also on our brains, manipulating our mental devices in order to bring out the appropriate (not understood and unintended) outcomes.*

The problem that needs to be solved is not just how a given *equilibrium* or state of *coherence* is achieved and a stable *order* emerges. In order to have a "social order" or an "institution", spontaneous emergence and equilibrium are not enough. The emergent structures must be "functional" (Sect. 10).

This problem appears in other sciences as the problem of "functions" (social and biological) – the question how certain effects of behaviors of anticipatory and intentional agents impinge on these agents and their "intentions" (see Castelfranchi 2001, and Sect. 10). Without a theory of emerging functions among cognitive agents, social behavior cannot be fully explained. No theory of social functions is tenable that does not solve this problem, first formulated by Adam Smith.

Adam Smith's original formulation of "THE problem" is – to me – much deeper and clearer than Hayek's formulation.

According to Smith, the great question is how "*(the individual) – that does neither, in general, intend to pursue the public interest, nor is aware of the fact that he is pursuing it, . . . is conduced by an invisible hand to* pursue *an* end *that is not among his* intentions".

Smith's "Invisible Hand" situation can be characterized as follows:

1. there are intentions and intentional behavior;
2. some unintended and unaware (long term or complex) effect emerges from this behavior;
3. but it is not just an effect, it is an *end* we "pursue", i.e. its orients and controls -in some way- our behavior: In some way, "operate *for*" that result.

Now, assuming this view is correct (as I believe), the problem posed by Smith is this:

- how is it possible that we *pursue* something that is not an intention of ours; that the behavior of an intentional and planning agent be goal-oriented, finalistic (*'end'*), without being intentional?
- in which sense the unintentional effect of our behavior is an "end" of the agent?

In sum, there cannot be a foundational severance between the social sciences and the cognitive sciences:

- a cognitive theory that ignores sociality (both interactive and collective, both historical and institutional) explains nothing about human mind and conduct, which are historically, institutionally and culturally determined and "written";
- a social science that is not founded on an adequate model of individual actors and their mental processes (what and why the actors believe, understand or do not understand, want or do not want) explains very little; it may propose "laws", but these laws do not describe the proximal causal mechanisms.

Besides, note that even if the only *raison d'être* of the social sciences would be the existence of complex and unintended social effects of individual actions, this justification of their necessity would not necessarily identify an exclusive object for them.[5]

In any case, the foundational task of the SSs is to give an account of the processes of emergence, given the mind and its contents, and conversely, an account of the processes of "immergence": How is it possible that societal objects, effects, and structures set up and establish themselves, and work without being understood and wanted? And how do they manage to get reproduced by feed-backing into the actors' minds? This, too, needs to become an intrinsic part of sociological theory. Which cognitive mediators are necessary for a given macro-phenomenon or structure (a political power structure, or a class or ethnic division) to emerge? What are the processes and powers (education, conformism, membership and identity, values, moral, religion) that construct and guarantee them?

## *1.1 The* **"Cognitive Mediators"** *of Social Phenomena*

As argued above, social and cultural phenomena cannot be fully understood without explaining how they arise by being represented in individual agents' minds (i.e., *through their mental "counterparts" or "mediators"*). As the social psychologist Kurt Lewin put it: "The most important fact concerning human interactions is that these events are *psychologically represented* in *each* of the participants" (Kurt Lewin 1935).

Lewin is certainly right, but ambiguous (and incomplete, see Sect. 1.2): Exactly in which sense are human interactions "psychologically represented"? Do social actors fully understand what they do? For this reason, I prefer the term "mediators" for these mental representations (Castelfranchi 2000): "Mediators" because they are mental states necessary for producing a given social phenomenon or structure, but *without (necessarily) being mental representations (an understanding or intending) of the social phenomena that are produced by the behaviors that they determine.*[6]

For example, one can play and reproduce a "social function" (being a father, consumer, the witness of a promise, "public opinion", the follower of a leader) without necessarily understanding this social function; nevertheless, one needs to have something specific to that function in one's mind to be able to reproduce it (Sects. 1.2 and 10).

[5] Can we be sure that without the emergent complexity of social phenomena we could make do without sociology, cultural anthropology, political science, etc.? Wouldn't these sciences still be necessary to understand collective intentional and organized behaviors, or to understand roles, institutional acts, norms, as well as values, trust, groupness, alliances, conflicts?

[6] This conceptualization obviously requires a richer cognitive model (architecture) for agents than that assumed in many formal and computational AI and ALife models, an agent architecture closer to those developed in psychology, cognitive science, and in cognitive approaches in economics, sociology and organization studies.

Social phenomena are caused by the behaviors of agents, but the behaviors of agents are caused by the *mental mechanisms* that control and (re)produce them.

For example: My Social Power lies in, indeed consists of, the others' *goals & beliefs* (see below); that's why we need mind-reading: Not for adjusting ourselves, but for manipulating and exploiting the others or for helping or punishing them.

As another example, norms exert their effects on behavior by working *through* the minds of the agents. But how exactly are norms "represented"? Which are the *proximate* mechanisms underlying normative behavior?

### *1.2 Mind Is Not Enough*

To explain what happens at the societal and collective levels, it is necessary to model the mind of the actors; however, "necessary" doesn't means "sufficient": The individualistic cognitive approach is not sufficient for the theory of social processes (even for just modeling interactions, joint and collective attitudes and actions; not yet even speaking of societies). The reason is that social actors do *not* understand, negotiate, and plan all their collective behavior and cooperative activity.

This is the real challenge non only for the behavioral and cognitive sciences but for MAS and Social AI and computer-supported societies: *Reconciling "Emergence" and "Cognition"*. Emergence and cognition are not incompatible with one another, nor are two alternative approaches to intelligence and cooperation. There are two reasons for this.

On the one hand, cognition itself has to be conceived as a level of emergence (from sub-symbolic to symbolic; from objective to subjective; from implicit to explicit). On the other hand, emergent and unaware functional social phenomena (e. g. emergent cooperation, swarm intelligence) do not exist only among sub-cognitive agents (Steels 1990; Mataric 1992), but also among intelligent agents. Therefore, even for a theory of cooperation and society among intelligent agents, mind is not enough (Conte and Castelfranchi 1996). It is very important to see the limits of deliberation and contracting in the production of complex social behavior: cognition alone does not explain social complexity (as Hayek noted).

In this paper, I will not deal with all the facets of this difficult problem; I will mainly present a *basic ontology of social concepts* (describing the most important atoms and some molecules of social life).

## 2 Some Preliminary Clarifications

Let me first shortly clarify some usual misuses of social and cognitive notions. As said conceptual clarification is a crucial job of any science.

## *2.1 Mental "Representations"*

"Cognitive or mental[7] representation" is frequently used as synonym for "knowledge" or "belief", of "doxastic representation". This is not correct: There are at least two kinds (uses, functions) of mental representations: doxastic representations (beliefs, knowledge, data, ...) and motivational representations, that is, "goals" (intentions, desires, projects, aims, objectives, preferences, ...). I particularly emphasize the importance of the latter, finalistic representations (goals), the representations that drive and orient the behavior and give meaning to it. Knowledge is actually just a resource, instrumental to goal pursuit and achievement: cognitive agents are purposive (goal-driven) agents. Sociality, too, is based on goal relationships.

In my view, economics, game theory, and even primatology and related sciences do not pay sufficient attention to the centrality of goals and their social dynamics: Goal-Adoption, Goal-Delegation, Goal-Induction etc. Beliefs are just there for the purpose of goals, since only Goals can control our behavior. Of course also beliefs are indispensable for a successful behavior, but I dispute their "centrality" in mind conception (Castelfranchi 2012a, b).

In particular, *mind reading* is not only there to allow understanding and predicting others' behavior and allow to coordinate one's own behavior to it; it is primarily there for generating appropriate Goals about the other's mind, for changing his behavior by *manipulating* his mind. Likewise, *norms* (and *values*) exist for influencing our behavior by changing our mind, our preferences and intentions (by changing our beliefs): they represent (other's and our own) goals about our goals (that proximately control our behavior).

## *2.2 "Cognitive" Is Not "Rational"*

Mind is not necessarily rational, and "cognition" is therefore not a synonym for rationality. Rationality is a special way of working of the cognitive apparatus: Cognition and action are rational if beliefs are well grounded in evidence, inferences are not biased by wishful thinking, illusions or delusions, decisions are based on these well-grounded beliefs and on a correct consideration of expected risks and advantages with their respective values. The rational mode is at best an idealized model of the workings of the human mind, and it can perhaps serve as a normative guideline; but it is not a model of how the mind actually works. Actual belief-formation and choice usually does not conform to this ideal model (by the way,

[7] "Cognition" and "mind" are clearly not synonyms for "consciousness". I will ignore the concept of consciousness, which covers on the one hand very different kinds of mental states, and on the other hand describes but a special state (and use) of mental representations.

only 10 % of human eyes conform to the ‘normal’ eye as described in texts of oftalmology).

The truth is, for example, that humans have a variety of different motives, all of which must in principle be considered to be able to explain their behavior. If we do this, many claims about the alleged “irrationality” of human subjects in economic or strategic decisions are revealed as unjustified, whereas the rational-decision making model is immediately revealed as being arbitrarily prescriptive, dealing only with one (“pleasure maximation”; “economic gain”) or a few presumably “rational motives” (sic!).

Although the variety of human motives can explain many presumed “deviations” from rationality, there also exist a variety of cognitive mechanisms (more or less “deliberative” or “reactive”) that govern behavior. Thus, it is also true that humans do not always follow rational decision making principles, and that other mechanisms (based on rules, routines, associations) must be modeled as well. The solution for obtaining a more adequate model of human decision making does not consist of simply equipping a rational decision-making model with some “emotional distortion” mechanism, or by bypassing the question of rational decision making by claims about how “rational” (adaptive) emotional impulses are. This juxtaposition of rationality and emotion is just a verbal solution. What is needed is an articulated model of the intertwining of explicit deliberation processes and emotional processes, and such an intertwining must be founded on a broader cognitive model of the processes on which both deliberation and emotion build. For these reasons, the typical economic approach to emotion – letting emotions contribute to the utility function while leaving the decision-making mechanism untouched – is insufficient, because too conservative. However, I will not discuss emotion in this paper (see Castelfranchi 2003c).

## 2.3 Social vs. Collective

The term “social action” is frequently used – in both the cognitive sciences and in philosophy – as the opposite of “individual action”, thus as denoting the action of a group or a team, rather than that of an individual. Social action, according to this understanding, is a form of collective activity, possibly coordinated and orchestrated, thus leading to joint action. However, we should not confuse or identify social action and social intelligence with collective action and collective intelligence.

Many theories about joint or group action (for example, Tuomela 1993; Tuomela and Miller 1988; Levesque et al. 1990), try to build group action up from individual actions: for example by reducing joint intentions to individual non-social intentions, joint plans to individual plans, group commitments (to a given joint intention and plan) to individual commitments to individual tasks. In my view, however, this is too simple because, in this analysis, a decisive intermediate level between individual and collective action is bypassed. Thereby, the real foundation of all sociality

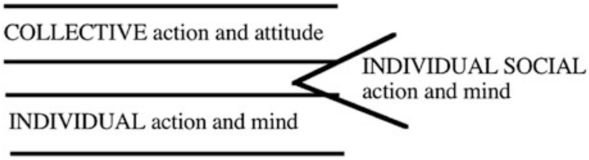

COLLECTIVE action and attitude
INDIVIDUAL SOCIAL action and mind
INDIVIDUAL action and mind

**Fig. 1** A frequently missed layer

(cooperation, competition, groups, organization, etc.) is missed: i.e. the individual social action and mind (Fig. 1).

*One cannot reduce or connect action at the collective level to action at the individual level if one passes by the* ***social*** *character of the individual action.* Collective agency presupposes individual *social* agents: the individual social mind is the necessary precondition for society (among cognitive agents). Thus we need a definition and a theory of *individual* social action and its forms (Castelfranchi 1997, 2000).

## *2.4 The Intentional Stance:* Mind Reading

Individual action is social or non social depending on the mind of the agent and on its purposive effects. The concept of social action cannot be a behavioral notion, i.e. one that is just based on an external description of behavior; it requires to model the mental states of agents and to consider agents' representations (both beliefs and goals) of the minds of other agents.

## *2.5 Social Action vs. Communication*

The notion of social action (that is foundational for the notion of agent or actor) cannot be reduced to communication, or modeled on the basis of communication. *Agents cannot be called "social" because they communicate*; it is the other way around: they communicate because they are social. They are social because they act in a common world where they interfere with, depend on, and influence each other.

## *2.6 Social Action and Communication vs. Cooperation*

Social interaction and communication are mainly based on some exercise of *power* (Castelfranchi 2003a), i.e. on either unilateral or bilateral attempts to influence the behavior of other agents by changing their minds. Both interaction and communication are frequently aimed at blocking, damaging, or aggressing against other agents, or at competing with them. Social interaction (including communication) is not necessarily the joint construction and execution of a multi-agent plan, of a shared script, necessarily based on mutual beliefs. It is not necessarily a cooperative activity (Castelfranchi 1992).

## 3 The Goal-Oriented Character of Agents and Actions

Sociality presupposes agents and *goals*. At a very basic level, an agent is any entity *able to act*, i.e. to produce some causal effect and some change in its environment. Of course this broad notion (including even natural forces and events) is not enough for sociality. We need a more complex level of agenthood: An agent is an entity that receives and exploits relevant information from and about the world. In which sense this is "information" for the agent? Why is it "relevant"? Our agent bases its action on it, i.e. on its perception of the world. In such a way its behavior or reaction is adapted to the environment; but on the other side (thanks to the functional action) the environment is adapted to and by the agent.

In other words, the agent's behavior is aimed at producing some result. In this case, we are talking about a *goal-oriented action* and a goal-oriented agent (McFarland 1983; Conte and Castelfranchi 1995).

Systems *oriented* towards some goal (although without any explicit internal representation of those goals) can be social, can exhibit social behavior. An "agent" can be helped or damaged, favored or threatened, it can compete or cooperate. These notions can meaningfully apply only to systems endowed with some (mental or functional) form of goal.

Among the goal-oriented systems we will consider in particular *goal-directed* system. In these systems, not only action is based on perception, it is based on the perception of the action's effects and results, and the agent regulates and controls its action on this basis. In other words, *the agent is endowed with "goals",* as *internal* anticipatory and regulatory representations of action results.

To be more precise, *actions* are teleonomic or goal-oriented behavior. We allow for goal-oriented behaviors that are not goal-directed (for example in many animals, or in functional artifacts), i.e. behaviors that are not motivated, monitored and guided by an internal (mental) representation of the effects.

A **goal** is a mental representation of a world state or process that is candidate for[8]:

– *controlling* and *guiding* action thanks to repeated tests of action's expected or effective results against the representation itself;

[8] I use "goal" as a general family term for all motivational representations, including desires, intentions, objectives, motives, needs, ambitions, concerns etc. Alternatives are "concerns" (Frijda 1986) or "desire" (Reisenzein 2009; Bratman 1990). However, "desire" – for me - is not a good general term, since (as used in common sense) it does not comprehend duties, obligations, needs, and other types of goal (Castelfranchi 2012a, b).

- determining the action search and selection;
- qualifying its success or failure.

This notion of goal-directed behavior is based on the very operational notion of goal and "purposive behavior" proposed by Rosenblueth and Wiener (1968), and developed, in psychology, by Miller et al. (1960). Unfortunately, this very clear definition of the purposive character of action is currently quite disregarded.

Action and social action is also possible at the reactive level, among sub-cognitive agents (like bees). By "sub-cognitive" agents I mean agents whose behavior is not regulated by an internal explicit representation of its purpose and by explicit beliefs. Sub-cognitive agents are for example simple neural-net agents, or mere reactive agents.

I will here analyze mainly goal-directed actions that require cognitive agents, i.e. agents whose actions are internally regulated by goals, and whose goals, decisions, and plans are based on beliefs. Both goals and beliefs are cognitive representations that can be internally generated and manipulated, and that can participate in inference and reasoning.

Since a goal-directed agent may have more than one goal that is active in the same situation, it must have some choice or decision mechanism; and this presupposes that the goals have a 'value', or 'importance' that allows to compare them and have preferences for them. Goal-directed agents also have an action repertoire (skills), some recipes, and some resources. However, their abilities and resources are limited, and therefore they are able to achieve only some of theirs goals.

I will say something on functions and their relations with intentions later (see Sect. 10).

## 4 From Non-social Action to Social Action: Beliefs & Goals About the Other's Mind

Any action is in fact inter-action, since its environment is never a passive entity: the action is aimed at producing effects *on* the environment and is controlled by the feedback *from* the environment. More than this: there is always some "delegation" to the environment of part of the causal process determining the intended effect, some reliance on the "activity" of the environment and of its causal forces. Hence, all actions are in fact interactions between agents and the environment. However, this does not imply that just any action should be called a "social" action. The environment is – as just said – a causal "agent" involved in the realization of our plans or actions, but this (inter)action need not be social, because the environment need not be, or include, a goal-oriented agent. For example, we can exploit the sun, but we cannot help the sun. Of course, if a primitive or superstitious person considers nature and natural objects as animate beings, from his subjective point of view, he is performing a social action (and collaboration) when he, for example, seeks the help of the "spirits" of the objects (e.g., by worshipping them).

Exploiting biological nature is social behavior at the weakest level, because the plants, ferments, viruses, etc. that we exploit or try to avoid or control (by preventing their activity) are in fact goal-oriented systems, and we treat them as such. While we cannot collaborate with sun and rain, since they do not have “ends”, plants are in some sense (unintentionally) collaborating with us, because they have “goals” such as producing fruits etc., and we consider in our plans not only their effects, but also their “goals” – we collaborate with them. Agriculture is in fact a kind of “collaboration” between humans and nature.

A social action in the full sense of the word is *an action that deals with another entity as an agent (in strict sense); i.e. as an active, autonomous, goal-oriented entity.*

For cognitive agents, a social action is an action that deals with another cognitive agent considered as a cognitive agent, i.e. an agent whose behavior is regulated by beliefs and goals. In social actions, the agent takes an intentional stance towards the other agents, i.e. it represents the other agent’s mind in intentional terms (Dennet 1981).

Consider a person (or a robot) running in a corridor who suddenly changes direction or stops because of a moving obstacle that crosses its path. Such a moving obstacle might be either a door opened by the wind, or another person or robot. The nature of the agent’s action (social or not) does not change depending on the nature of the obstacle: If x acts towards another agent as if it were just a physical object, her action is not a social action. Whether it is a social action or not depends on how x *subjectively* represents the other entity in her action plan. Hence, *an action related to another agent is not necessarily social*. The opposite is true as well: A purely practical action that does not involve other agents, such as closing a door, may be or become social. Closing the door would be social, for example, if we perform this action to prevent other agents from entering or looking inside the room, whereas the same action performed to block wind or rain or noise is not social. Hence, *not behavioral differences but goals distinguish social from non-social action*.

We may call “weak social action” social action that is based just on *social beliefs*: beliefs about other agents’ minds or actions; and “strong SA” actions which are also directed by *social goals*.

The true basis of any level of SA among cognitive agents is *mind-reading* (Baron-Cohen 1995): the representation of the mind of the other agent.

Notice that beliefs about the other’s mind are not only the result of verbal and nonverbal communication about mental states, or of scripts and roles (Castelfranchi 2012a, b), or stereotypical ascription; the result also of interpretation of the other’s behavior. In a cognitive agent, the other’s behavior necessarily becomes a “sign” of his mind. This *understanding*, as well as behavioral and implicit communication is, before explicit communication (special message sending), the basis of reciprocal coordination and collaboration.

But we do not only have beliefs about the other’s mind, we also have goals about her behavior and thus her mind, i.e. the beliefs and goals regulating her behavior. We act in order to change the other’s mind and behavior; to manipulate or exploit the other, or else to help and promote him (Sect. 7).

But why? Why do we want to elicit a particular behavior or reaction from another agent? Part of the answer is: We do this because we need others – their skills and resources or mental responses – for our goal-achievement: We depend on them (even for helping them).

## 5 Dependence, the Reason of Sociality

The real structural basis and origin of sociality is Dependence and Power, but Dependence and Power presuppose goals and are "mentally grounded"; they depend on the minds of the interacting Agents: not only in the sense that they have (different) goals and competences/skills, but because of their beliefs and knowledge about their relations; the "cognitive emergence" (see later).

Social Dependence is due to being in a "common world" that is to "interference": My actions can facilitate or prevent your goal-achievement, and/or vice versa.

> *X Depends on Y as for a given action/resource (a) of Y relatively/for a given goal (that p).* (Castelfranchi et al. 1992; Sichman et al. 1998).

Dependence is first of all an *objective social relation*: the combination of lack of power of one agent (relative to one of its own goals) and of the corresponding power of the other agent. But of course the perception of dependence ("subjective dependence") is crucial (Fig. 2).

The dependence network *determines* and *predicts* partnerships and coalitions formation, competition, cooperation, exchange, conflicts, the functional structure in organizations, rational and effective communication, and negotiation power, power over the other, etc.

Several typical, important dependence patterns can be distinguished (such as transitive, reciprocal, mutual, OR, AND dependence). Dependence also has a quantitive aspect, it differs in degree: Other factors constant, X becomes more dependent on Y the higher the value and number of the goals of X whose fulfillment requires the cooperation of Y, and the fewer alternatives to Y (competitors) exist.

Given our inter-dependence, we can "compete" and "fight"; or/and we can "exchange" and "cooperate" (we can also do both). Both directions (solidarity and "homo homini lupus") emerge spontaneously and are later orchestrated, and organize social action and society.

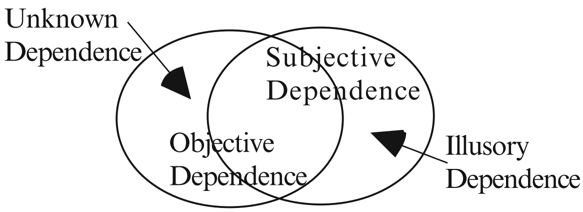


**Fig. 2** Mind about dependence relations

The main function of pro-social or positive sociality is *the multiplication of the power* of the participating agents.[9] Hence, different from Hogg and Huberman (1992), I do not assume that the greatest advantage of (cooperative) sociality is to speed up the search for solutions to *common* problems, or to find better solutions to them, but rather to *multiply individual powers*: Any agent, while remaining limited in its capabilities, skills and resources, finds the number of goals it can pursue and achieve increased by virtue of its "use" of others' skills and resources. In a sense, any agent's power limits, and its differences from others in the kind of power it is endowed with, turn into an advantage (this is Durkheim's perspective): Although not omnipotent, *the agent is able to overcome its cognitive, and practical limits through* "*sociality*".

However, within this general phenomenon we need to distinguish between two very different kinds of power improvement: the "circulation of powers", and "complex power construction" (Castelfranchi 2011). As to power networks and minds see Sect. 10.

## 6 Relying on (Delegating): *Making the Other Realize Our Goal*

I will now examine those elementary forms of social actions that are the basic ingredients of helping, exchange, cooperation, and then of partnership, group and team work. Let us consider them in their "statu nascenti", starting from the simple unilateral case.

On the one side, there is the mental state and the role of the future "client" (who relies on another agent's action to achieve her goal) -let us call this *Delegation* or *Reliance*. On the other side, there is the mental state and role of the future "contractor" (who decides to do something useful for another agent, by adopting a goal of hers) – let us call this *Goal Adoption*.

In **Delegation** *x needs or likes an action of y and includes it in her own plan: she relies on y. She plans to achieve p through the activity of y.* So, she is constructing a MA plan and *y* has a share in this plan: *y*'s delegated *task* is either a state-goal or an action-goal (Castelfranchi and Falcone 1997; Lorini et al. 2007).

**Unilateral Weak Delegation**

In Unilateral Delegation there is neither bilateral awareness of the delegation, nor agreement: *y* is not aware of the fact that *x* is exploiting his action. One can even "delegate" some task to an object or *tool*, relying on it for some support and result

[9] It seems that the less the "individual Self-Sufficiency" (the number of self-realizable goals) is, the more sociality becomes useful as a multiplier of power. (However, the function is complex, because we need agents with high "power of" (capability, resources), and low "Self-Sufficiency"). In other terms, the more the individuals are dependent on each other, the more sociality multiplies their power. This is one of the reasons why division of labor and specialization are so productive.

(Conte and Castelfranchi 1995, chapter 8). In the weakest and passive form of unilateral delegation *x* is just exploiting the autonomous behavior of *y*; she does not cause or elicit it.

As an example of weak and passive, but already social delegation (which is the simplest form of social delegation) consider a hunter who is ready to shoot an arrow at a flying bird. In his plan the hunter includes an action of the bird: to continue to fly in the same direction (which is a goal-oriented behavior); in fact, this is why he is not pointing at the bird but at where the bird will be in a second. He is delegating to the bird an action in his plan; and the bird is unconsciously and unintentionally "collaborating" with the hunter's plan.

**Delegation by Induction**
In this stronger form of delegation *x* is herself eliciting or inducing *y*'s behavior in order to exploit it. Depending on the reactive or deliberative character of *y*, the induction is either based on some stimulus or on beliefs and complex types of influence.

As an example of unilateral Delegation by induction consider a fisherman: unlike the hunter, the fisherman elicits by himself -with the bait- the fish's action (snapping) that is part of his plan. He delegates this action to the fish (he does not personally attach the fish to the hook) but he also induces this reactive behavior.

**Delegation by Acceptance (Strong Delegation)**
This Delegation is based on *y*'s awareness of *x*'s intention to exploit his action; normally it is based on *y*'s adopting *x*'s goal (social goal-adoption), possibly after some negotiation (request, offer, etc.) concluded by some agreement and social commitment. *X* asks *y* to do what she needs and *y* accepts to adopt *x*'s goal (for any reason: love, reciprocation, common interest, etc.). Thus to fully understand this important and more social form of Delegation (based on social goals) we need a notion of social goal-adoption (see Sect. 6); we have to characterize not only the mind of the delegating agent but also that of the delegated one.

## 6.1 Levels of Delegation

Given a goal and a plan (sub-goals) to achieve it, *x* can delegate goals/actions (tasks) at different levels of abstraction and specification (Falcone and Castelfranchi 1997). We can distinguish between several levels, but the most important are the following ones:

- *pure executive delegation vs. open delegation;*
- *domain task delegation vs. planning and control task delegation (meta-actions)*

The object of delegation can be minimally specified (*open delegation*), completely specified (*closed delegation*) or specified at any intermediate level.

*Open delegation* necessarily implies the delegation of some meta-action (planning, decision, etc.); it exploits intelligence, information, and expertise of the delegated agent. Only *cognitive delegation* (the delegation of a *goal*, an abstract action or plan that need to be autonomously specified) can be "open": thus, it is *something that non-cognitive agents cannot do.*

### 6.1.1 Necessity and Advantages of 'Open Delegation' in Collaboration

It is worth stressing that *open delegation* is not only due to *x*'s preferences, practical ignorance or limited ability. It can be also due to *x*'s ignorance about the world and its dynamics: *Fully specifying a task is often impossible or inconvenient,* because some local and updated knowledge is needed for that part of the plan to be successfully performed. Open delegation ensures the *flexibility* of distributed and MA plans.

The distributed character of the MA plans derives from the *open delegation*. In fact, *x* can delegate to *y* either an entire plan or some part of it (*partial delegation*). The combination of the *partial delegation* (where *y* might ignore the other parts of the plan) and of the *open delegation* (where *x* might ignore the sub-plan chosen and developed by *y*) creates the possibility that *x* and *y* (or *y* and *z*, both delegated by *x*) collaborate in a plan that *they do not share* and that *nobody* entirely knows: that is a *distributed plan* (Grosz and Kraus 1996). However, for each part of the plan there will be at least one agent that knows it. This is also the basis for Orchestrated cooperation (a boss deciding about a general plan), but it is not enough for the emergence of functional and unaware cooperation among planning agents (Castelfranchi and Conte 1992).

## *6.2 Motivation for Delegation*

Why should an agent delegate some action to another, trust it and bet on it? As we said, delegation is due to *dependence: X*, in order to achieve some goal that she is not able to achieve alone – be it a concrete domain action or result, or a goal like saving time, effort, resources – delegates the action to another agent both able and willing to do it. *X* either lacks some know how, or ability, or resource, or right and permission, and is depending on the other agent for them.

Of course, *X* can delegate actions that she is able to do alone; she just *prefers* to let the others perform them on her behalf. However, if *X* prefers exploiting the action of the other agent for her goal that p, this means that this choice is better to her, i.e. there is some additional goal she achieves by delegating (ex. saving effort, time, resources; or having a proper realization of the goal, etc.). Relative to this more global or complete goal which includes p, *X strictly* depends on the other. So the dependence relative to global intended results of delegation is the general basis of delegation.

## 7 Goals About the Other's Action/Goal

In Delegation *x has the goal that y does a given action* (that x needs to be done, and includes in her plan). If *y* is a cognitive agent, *x has also the goal that y has the goal* (more precisely *intends*) to perform that action. Let us call this "cognitive delegation", that is delegation to an intentional agent. This goal of *x* is the motive for *influencing y* (Pörn 1989; Castelfranchi 2003a), but it does not necessarily lead to influencing *y*. Our goals may also be realized by the independent evolution of the environment, including events and other agents' actions. Thus, it might be that *x* has nothing to do because *y* independently intends to perform the needed action.

Strong social action is characterized by *social goals*. A social goal is defined as a goal that is *directed toward* another agent, i.e. whose intended results include another agent as a cognitive agent: *a social goal is a goal about other agents' minds or actions*. Examples of typical social goals (strong SAs) are: changing the other mind, communication, hostility (blocking the goal of the other), cognitive delegation, adoption (favoring the goal of the other).

We do not only have *beliefs* about others' beliefs or goals (weak social action) but also *goals* about the mind of the other: *x* wants that *y* believes something; *x* wants that *y* wants something. We cannot understand social interaction or collaboration, nor organizations, without considering these social *goals*. Personal intentions of performing one's own tasks, plus beliefs (although mutual) about others' intentions are not enough.

For a cognitive autonomous agent to acquire a new goal, he needs to acquire some new *belief* (Castelfranchi and Paglieri 2007). Therefore, cognitive influencing consists of providing the addressee with information that is pretended to be relevant for some of his goals, and this is done in order to ensure that the recipient has a new goal.

### *7.1 Influencing Goal, Power, and Incentive Engineering*

The basic problem of social life among cognitive agents lies beyond mere coordination: *How to change the mind of the other agent? How to induce the other to believe and even to want something*? How to obtain that *y* does or does not something? Of course, normally -but not necessarily- by communicating to the other.

However, communication can only inform the other about our goals and beliefs (about his action): but *why should the other care about our goals and expectations?* Thus, in order to induce him to do or not to do something we need *power over* him, *power of influencing* him. His benevolence towards us is just one of the possible basis of our power of influencing him (authority, sympathy, etc. maybe others). However, the most important basis of our power over another agent is the fact that our actions, too, are potentially interfering with his goals: we might either damage or favor him, he is depending on us for some of his goals. We can exploit this (his dependence, our reward or incentive power) to change his mind and induce him to do or not to do something (Castelfranchi 2003a).

*Incentive engineering*, i.e. manipulating the other's utility function (his outcomes or rewards due to goals achievement or frustration), is not the only way we may have to change the mind (behavior) of the other agent. In fact, in a cognitive agent, pursuing or abandoning a goal does not depend only on preferences and on beliefs about utility. To pursue or abandon his intention, *y* also needs to have a host of beliefs that are not reducible nor related to his outcomes. For example, to do p *y* must believe that p is possible, that he is able to do p, that p's preconditions hold, that necessary resources are available, etc. It is sufficient for *y* that *x* modifies one of these beliefs in order to induce *y* to drop his intention to p and to pursue some other goal.

Thus, the general formula of influencing cognitive agents' behavior is not incentive engineering, but *"modifying the beliefs which support goals and intentions and provide reasons for behavior"*. Modifying beliefs about incentives represent only a sub-case (Castelfranchi and Paglieri 2007).

## 8 Social Goal Adoption: To Act in Order to Realize a Goal of the Other

Let us now look at social action from *y*'s (the contractor or the helper) perspective. Social goal-adoption (shortly *G-Adoption*) deserves a more detailed treatment, because: (a) it is the true essence of all pro-social behavior, and has several different forms and motivations; (b) its role in cooperation is often not well understood. In fact, in most existing analyses, agents are either just assumed to have the same goal, or the adoption of the goal from their partners is not explicitly accounted for (Tuomela 1993; Tuomela and Miller 1988; Levesque et al. 1990); or the reasons for adopting the other's goal and for taking part in the collective activity are not explored.

*In* **G-Adoption** *y is changing his mind: he comes to have a new goal, or at least to have new reasons for an already existing goal*. The reason for this (new) goal is the fact that another agent *x* wants to achieve this goal: *y* comes to know this, and as a consequence decides to make/let her achieve it. So, *y* comes to have the same goal of *y*, *because* he knows that it is *x*'s goal.

However, this characterization of G-adoption is too broad: this social attitude and action should not be mixed up with simple *imitation*, which might be covered by that definition. In G-adoption, *y* has the goal that p (wants p to be true) because he wants *x* to achieve it. In other words, *y is adopting a goal of x's if y wants x to obtain it and as long as y believes that y wants to achieve that goal* (Conte and Castelfranchi 1995).

### *8.1 Goal-Adhesion or Compliance*

Among the various forms of G-adoption, *G-adhesion* or *Compliance* has a special relevance, especially for modeling agreement, contract and team work. Compliance occurs when the G-adoption is due to the other's (implicit or explicit) request. It is the opposite of spontaneous forms of G-adoption.

In compliance, not only has *x* has a goal p that *y* adopts, but *x* herself has the goal that *y* does something to achieve p, and (as a consequence) also the goal of letting *y* know about her wish (request). Thus, in G-adhesion, *y* adopts *x*'s goal that he adopts her goal, i.e. he complies with *x*'s wishes.

In order to satisfy *x*, not only *y* must achieve p (like in spontaneous and weak G-adoption), but he must also let *x* know that he performed the expected and delegated action and produced p.

G-adhesion is the strongest form of G-adoption. Strong delegation is required for adhesion. G-adhesion is the basis of all agreements, negotiations, speech acts, norms, etc. that are based on the communication by *x* of her intention that the other does something, or better adopts her goal (for example obeys). In all these cases, G-adhesion is what really matters.

## *8.2 Endogenous and Exogenous Goal-Sources*

Through social *goal-adoption* we obtain a very important result as for the architecture of a social agent:

- Goals (and then intentions) do not all originate from desires or wishes, they are not all derived from internal motives. A social agent is also able to "receive" goals from the outside: from other agents or from the group; as requests, needs, commands and norms.

In architectural terms, this means that there is not a single origin of potential intentions or candidate goals. Rather, there are several origins or sources of goals: bodily needs; goals activated by beliefs; goals elicited by emotions; goals generated by practical reasoning and planning; and goals adopted (i.e. introjected) from the outside. However, all of these goals presumably have to converge at some level or stage in the same processing mechanism, to become intentions and hence to be pursued by actions.

## *8.3 Motivation for G-Adoption*

The adoption of goals from others does not coincide with *benevolence*. However, a relation of benevolence to another agent is indeed a form of generalized adoption and contributes to the motivation for G-Adoption.

Benevolence (pity, altruism, love, friendship) for another is the basis of a *terminal* (noninstrumental) form of G-adoption. However, Goal-adoption can also occur because it is seen as *instrumental to the achievement of selfish goals*. For example, feeding chickens (satisfying their need for food) is a means for eventually eating them. Instrumental G-Adoption also occurs in social “exchange” in strict sense: “do ut des” (reciprocal, conditional G-Adoption).[10]

Another motive-based type of G-Adoption (that might be considered a sub-type of instrumental goal adoption) is *cooperative* G-Adoption: *y* adopts *x*’s goal because he is co-interested in (some of) *x*’s intended results: they have a common goal.[11] Collaborative coordination is just one example of this.

The distinction between these three forms of G-Adoption is quite important, because their different motivational bases allow important predictions on *y*’s “cooperative” behavior. For example, if *y* is a rational agent, in mere “exchange” he is interested in and should try to cheat, not reciprocating *x*’s adoption. On the contrary, in “cooperative” adoption in strict sense *y* normally is not interested in free-riding, because he has the same goal as *x* and they are *mutually dependent* on each other as for this goal p: both *x*’s action and *y*’s action are necessary for p, so *y*’s defeating *x* would be self-defeating. Analogously, while in terminal and in cooperative adoption it might be rational in many cases to inform *x* about difficulties, obstacles, or defections (Levesque et al. 1990), in *exchange*, and especially in forced, coercive G-Adoption, this is not the case at all.

## *8.4 Levels of Collaboration*

In analogy to delegation, several dimensions of adoption can be characterized (Falcone and Castelfranchi 1997). In particular, the following levels of adoption of a delegated task can be considered:

- *Literal help*: *x* adopts exactly what was delegated by *y* (elementary or complex action, etc.).
- *Overhelp*: *x* goes beyond what was delegated by *y,* without changing *y*’s plan.
- *Critical help*: *x* satisfies the relevant results of the requested plan/action, but modifies it.
- *Overcritical help*: *x* realizes an Overhelp by, at the same time, modifying or changing the plan/action.
- *Hyper-critical help*: *x* adopts goals or interests of *y* that *y* himself did not consider or did not delegate to him; by doing so, *x* does not perform the action/plan, nor reach the results he was asked to reach (but in the interest of *y*).

[10] A bilateral dependence relation between to merely selfish guys, with their own personal goal; As definitely characterized by Adam Smith: “*It is not from the benevolence of the butcher, the brewer, or the baker that we expect our dinner, but from their regard to their own interest. We address ourselves, not to their humanity but to their self-love, and never talk to them of our own necessities but of their advantages*”. This is market and “exchange” in strict sense.

[11] This is for us “cooperation” *in strict sense* (not covering for example mere “exchanges”). We need each other but not for our own independent results (goals) but just for one and the same result, objective (at least at a given layer).

On the basis of these distinctions, one can characterize the *level of collaboration* of the adopting agent. An agent that helps another by doing just what is literally requested to do, is not a very collaborative agent. He has no initiative, he does not care for our interests, does not use his knowledge and intelligence to correct our plans and requests that might be incomplete, wrong or self-defeating.

A truly helpful agent should care for our goals and interests, and beyond our explicit delegation or request. However, *only cognitive agents can non-accidentally help beyond delegation*, recognizing our current needs in each case.

Of course, there is also a danger involved in taking the initiative of helping us beyond our request. Troubles may be due either to misunderstandings and wrong ascriptions or to conflicts and paternalism.

## *8.5 Altruistic Acts or Minds?*

To regard a given act as "altruistic" requires to "judge the agent's intentions". "Altruistic" is a *subjective* notion; it just depends on the mental representations (in particular the motivational ones) *ascribed to* the agent and underlying his act[12]; it is not – if applied to cognitive agents – an objective and behavioral notion. It is not sufficient for altruism that X's behavior is (not accidentally, but functionally and regularly) beneficial for Y and expensive for X. It is not even enough that X's behavior is intended to benefit Y. Rather, altruism is a matter of final motives, of the ends of the act.

In our view (Lorini et al. 2005) it is impossible to solve the problem of the existence or not of "true" altruistic actions and people without making clear two issues:

(a) Being an "autonomous" agent, endogenously motivated and regulated by one's own "goals" (like in "purposive systems") is not the same as being "selfish". What common sense means by "selfish" or "egoist", is *not* that one is driven by "one's own" internal motives and choices; and "altruist" does *not* mean that one is hetero-regulated. What these terms refer to is the nature and origin of the regulating goals; but those goals are always the agent's own goals and preferences (Sober and Wilson 1998; Castelfranchi 2012a, b).

(b) As very clearly explained by Seneca[13] it is crucial to distinguish between expected positive results (the prediction of positive outcomes of my action)

[12] I mean that, if we consider X's act as truly altruistic we are attributing to X a specific motivational asset.

[13] "Itaque erras cum interrogas quid sit illud propter quod uirtutem petam; quaeris enim aliquid supra summum. Interrogas quid petam ex uirtute? ipsam. Nihil enim habet melius [enim], ipsa pretium sui. An hoc parum magnum est? Cum tibi dicam 'summum bonum est infragilis animi rigor et prouidentia et sublimitas et sanitas et libertas et concordia et decor', aliquid etiamnunc exigis maius ad quod ista referantur? Quid mihi uoluptatem nominas? hominis bonum quaero, non uentris, qui pecudibus ac beluis laxior est." Seneca, *De vita beata*, IX (http://www.thelatinlibrary.com/sen/sen.vita.shtml)

> and what "motivates" my action. Not all the expected positive outcomes of my action are *motivating* me; in other words, it is false that I act "in order to" achieve them, just because I predict them. In our view for *motivating* my action they should be *necessary* and *sufficient* conditions for my decision to act.[14]

Without this distinction and sophisticated modeling of motivations and goal processing we have just to be satisfied by "pseudoaltruism" (on this notion see Batson 1991) where the expected (at least internal) positive rewards of one's behavior are unduly identified with one's motivations.

Seneca's solution is very simple and intuitive (although we still do not have the corresponding psychological model!).

## 9 Joint Action: Two Cognitive Aspects

I will not deeply analyze the structure of joint and of collective actions, mental states, interests, emotions. Let me just – to be coherent with the very "basic" ontology of sociality I have focused on – just stress two issues: *the relevance of "social goals" for joint actions;* and the issue that *joint plans and intentions do not presuppose a fully shared mental representation* of what is jointly produced, even at the interpersonal or group level, not only at the societal or market level.

### *9.1 Social Goals as the Glue of Joint Action*

Although clearly distinct from each other, *social* actions and goals and *joint* actions and goals, are not two independent phenomena. A theory of joint actions, like a theory of groups and organizations, presupposes a theory of social goals and actions.

["But," says our adversary, "you yourself only practise virtue because you hope to obtain some pleasure from it." In the first place, even though virtue may afford us pleasure, still we do not seek after her on that account: for she does not bestow this, but bestows this to boot, nor is this the end for which she labours, but her labour wins this also, although it be directed to another end. As in a tilled-field, when ploughed for corn, some flowers are found amongst it, and yet, though these posies may charm the eye, all this labour was not spent in order to produce them—the man who sowed the field had another object in view, he gained this over and above it—so pleasure is not the reward or the cause of virtue, but comes in addition to it; nor do we choose virtue because she gives us pleasure, but she gives us pleasure also if we choose her.] (Of a Happy Life*, translated by Aubrey Stewart* From the Bohns Classical Library Edition of *L. Annaeus Seneca, Minor Dialogs Together with the Dialog* "*On Clemency*"; George Bell and Sons, London, 1900).

[14] This is the stronger condition. However, we have also broader and weaker cases: Where the expected positive outcome is just "necessary" for my decision but not "sufficient" (I need additional expected outcomes, given for example the costs or the risks); or where the expected positive outcome was "sufficient" for doing that action, but not "necessary"; since I would have done it also for other reasons and effects.

In fact, *social goals* in the minds of the group members are the real glue of joint activity.

As argued before, *if a collective goal is derived from or implemented in individual goals* (i.e., is not a primitive goal, in Searle's (1990) sense), *it necessarily implies some goal-adoption or interest-adoption*. Therefore, a collective goal can be reduced or at least analyzed – and unless it is a primitive goal, it *must* be analyzed – into individual goals by means of complementary *social* goals of delegation and of adoption in the minds of the individuals.

In particular, one cannot understand what really glues together a group or team, the group members' goals of influencing the others; the collaborative coordination, the commitments, the obligations and rights that relate one group member to the other. I cannot examine here the rather complex structure of a team activity, or collaboration, and the mental structures of the involved agents that required to make it possible; nor can I analyze here the mind of the group considered as a complex agent. Advanced formal descriptions of these phenomena are available elsewhere (Tuomela and Miller 1988; Levesque et al. 1990; Rao et al. 1992; Grosz and Kraus 1996; Tummolini 2010; Tummolini and Castelfranchi 2006). However, I would like to stress that individual social action and goals, as previously characterized, also play a crucial role in joint action (Castelfranchi 2003b). That is, no group activity, no joint plan, no true collaboration can be established without:

(a) the goal or better the expectation[15] of *x* (member or group) that *y* will intend to perform a given action/task p (*reliance/delegation/trust*);
(b) *x*'s "intention that" (Grosz and Kraus 1996) *y* is able and has the opportunity to do p; and in general the "collaborative coordination" of *x* relative to *y*'s task. This is derived from the *delegation* and from reliance and coordination necessary among actions in any plan.
(c) the *social commitment* of *y* to *x* regarding *p*, which is a form of goal-adoption or better adhesion.

Both *Goal-adoption* in collaboration and groups, and the *goal about the intention of the other* (influencing) are either ignored or just implicitly presupposed in the above mentioned approaches to group collective intentionality. They mainly rely on the agents' *beliefs* about others' intentions; i.e. a weak form of social action and mind. The same is true for the notion of cooperation in Game Theory.

As for the social commitment, it has been frequently confused with the individual (non social) commitment of the agent to his task (Castelfranchi 1995a, b).

[15] Not a simple prediction (a belief about a future state or event) but the combination of a belief about the future and a (convergent or opposite) goal.

### *9.2 Non-shared Plans in Collective Action*

The classical approach to collaboration attempted to analyze joint intentions and collaboration in terms of fully “shared” plans. This approach turned out to be unworkable; humans very often (even typically) work – intentionally and with mutual knowledge – on the execution of one and the same plan without representing in their minds those parts of the plan for which others are in charge. Our mental representations in these cases are complementary and partially blind. Even more radically ignorant and incomplete are our representations of the social ends in public and institutional collectives plans; for example, in norm-regulated public goods.

*Norms* not only work thanks to our only partial or complete blindness to their social aims, but in a sense they even require and “prescribe” this blindness: ideally we are expected simply to “obey”, not to agree, to negotiate, to do for our convenience, or intending the “common good”; we are not supposed to understand the *end* of the norm. The aim of the norms is that I trust that they are for the “common good” and for the Polis (not for private advantages), but not that I – by understanding what they are for – “cooperate” with the society.

I have to “submit” to the (group) authority: the decision about what to do in a particular situation has already been taken, and it was up to somebody else, not to me. By obeying I cooperate with a societal end that I need not to understand/know and that I have to “pursue” (Fig. 3). This is true (that I have to cooperate and that I in fact “cooperate” and “pursue” that end) even if I understand that end and disagree about it, as (to me) wrong or unfair. However, of course, differently from mere “social functions” there is a basic part of the norms that must be mentally represented and understood: A norm works as a norm only if it is recognized by us, and mentally represented as a norm – e.g., as a command from an entitled authority. A norm is not a *personal* request, claim, or imposition. To engage in norm-guided action – to act for that reason, for a sense of duty, for submission – we have to acknowledge the norm and the authority that stands behind it. Hence, norms impose and presuppose a partial understanding of the social artifact.

Society works thanks our *partial* understanding. We cooperate and jointly act thanks to the fact that we do *not* (fully) understand and intend what we are jointly doing.

## 10 Networks of Minds

A recent trend in sociology and technology is to focus on the modeling of the “networks”, the multi-agent “structure” of society, while bypassing the cognitive process of individual actors. Nets are indeed very important structures to study because of their specific effects, which they owe to their specific topology, connections, transmission of information etc., but Nets cannot replace minds, neither in theory nor in practice. What we really are dealing with in social science are networks of minds, and also a new layer of complexity.

Let me give just two examples of how crucial and basic constructs of the social sciences cannot not be founded in *cognitive*-pragmatic theories at the individual and collective layers. Behaviors, interactions, relations and their "network" structure do not replace minds and cannot do without minds. Also networks are networks of minds, and this gives rise to a specific emergence dynamics (cognitive emergence, immergence, ...) and new layers. I will look at two concepts whose centrality for the social sciences is beyond doubt, the concepts of "trust" and of "power".

### *10.1 Power Net*

What follows are just a few assertive and synthetic claims about power nets. I will focus on the basic interpersonal and collective levels and ignore the institutional or political ones.[16] However, of course already at these levels, power is grounded in the beliefs of the actors, the concrete behaviors caused by them, and the effects of the behaviors and their effects.

1. *I do not really have the "Power-of" doing, obtain something, if I'm not* <u>*aware*</u> *of such a Power of mine*: my skills, accessible resources, rights, competence, ...

   Human being are handicapped under this respect. As mentioned earlier, in order to intentionally pursue and do anything (i.e., formulate and perform an intentional action), we have to believe a lot of things; that the goal has not already been realized, that is up to us, that it is realizable and that we are able to realize it. If X is able to perform an action but is not aware of this fact, or if he does not believe that he can perform an action p (is able/skilled/competent and "in proper working condition") he will renounce, or not even formulate the intention to perform p. At his best he can decide to "try" to see if he is able to perform p, or whether p is possible or what the conditions for p are (Lorini et al. 2006).

   If I do not know that I have a given power, I cannot really (that is, intentionally and for my purposes) "exercise" it.
2. *I do not have a social Power-over Y if I do not know what Y wants, believes*, and thus if and how Y depends on me, and hence how I can interfere with his goal-achievement. As mentioned before, "Theory of Mind" is mainly for that; not – as usually said – for predicting Y's behavior and adjusting to it. It is for having Power-over the others and using it for "influencing", manipulating them.
3. *I do not have the Power-of-Influencing Y if Y doesn't know (*or rather, *believe!) that I have a Power-over him, and that I know that I have.*
4. *I do not have the needed "positioning" power in a Net*, a "comparatively" better evaluation (which is the basis of my being a "preferred partner" for exchange or cooperation) if the actors in the net do not have in their minds

[16] For a more complete analysis see Castelfranchi 2003a, 2011.

some representation of the various dependence relations in the net ("cognitive emergence").

5. *If Z doesn't know that I have the Power-of-Influencing Y, and can induce Y to harm Z, I have the Power-of-Infl Y to harm Z, but I do not acquire the Power-of-Infl Z by threatening Z with Y's possible harm.*

And so on. Beliefs about power give power, and build more complex layers and dynamics of power; *beliefs about the Net restructure the Net.*

## *10.2 Trust Net*

Trust is a psychological object, a mental state with specific affective and/or doxastic and motivational aspects: this much is obvious, and there are no serious disputes about it. However, a lot of "social" studies of trust relations and societal dynamics believe that one can put aside the "psychological" dimensions assumed as personal variations, personalities, biases, etc. (psychologisms!). This is not at all the case. The social *structuring* and effects of trust are cognitively grounded and must cognitively be explained.

On the one hand, social relations, interactions, and structures are explained by their mental foundations (what the individually actors assume and want); on the other hand, it is fundamental also at the collective level and for the macro-dynamics, the mental emergence of the trust relations, they representation.

Let us consider *a portion of a T net around X*: Y trusts X (for A); W trusts X (for A); Z trusts X (for A); Q trusts X (for A); now:

1. *What happens if X has such Net representation, if he knows about it?*

   He might for example exploit this multiplicity of T relations for a better negotiation power, for increasing the "price" of his service; or he might propagate this knowledge in his social network to gain esteem and reputation (and increase his perceived trustworthiness).
2. *What happens if Y knows that, compared with the situation where she just knows about her own trusting X?*

   Y might perceive X to have a better negotiation position, and thus – for example – decrease her request (price) or search for another partner; or Y might increase her T in X.

   Hence, the knowledge of the net feeds back on the trust relations and changes them.
3. *Everybody (Y, Z, W, Q, . . . ) knows about the Net, but independently (they don't know whether the others know about it or not).* As a consequence, the effects described in (2) are generalized, propagated. If Y assumes that (perhaps) the others do not have a full understanding of the trust net, he might try to deceive them: for example, he might tell them that doesn't trust X at all, in order to increase her own positioning, and to reduce X's opportunities.

4. *Everybody (Y, Z, W, Q, ...) knows about the Net, but also everybody knows that everybody knows.* This is the basis of genuine "collective" trusting X, which is also explicitly self-confirming, and can give rise to special moves (for example against the "monopolization" of trust).

## 11 Functions and Minds

As we saw, in modeling social phenomena we need mind, but we need also modeling *emergence* of complex blind and unplanned social structures, and of non-orchestrated cooperation.

Emergent intelligence and cooperation do not pertain only to reactive agents. Mind, too, cannot understand, predict, and dominate all the global and compound effects of actions at the collective level. Some of these effects are positive, self-reinforcing and self-organizing. There are forms of cooperation that are not based on knowledge, mutual beliefs, reasoning and constructed social structure and agreements. To be clearer: When I have an intention, I necessarily "intend" (have the decided goal of) doing something, but I also necessarily intend/want some result of the action, a further, higher goal. The problem is: *where does our intending stop along the goal chain?* Since the causal chain of effects (and of expected effects) continues far beyond what I explicitly intend, *what is my goal horizon.*

Not only does an actor not intend (consciously or unconsciously), nor even have in mind, all the consequences of his action, he does not even intend, or have in mind, all the "goals" of his actions! There are mental "goals", which are just in the heads of the others he is collaborating with or obeying to; and there are non-mental "goals" – mere "functions" of his behaviors, not only of the automatic or routine ones but also of those internally regulated by some personal goal. In *x*'s mind there are true goals, but they are only a part of the finalistc chain; the higher "goals" aren't represented in any mind, are not true goals, but just "functions" of that intentional behavior.

However, what kind/notion of *emergence* do we need to model these forms of social behavior? The notion of emergence simply relative to an observer or a merely *accidental* cooperation, are not enough for social theory and for artificial social systems. We need an emerging structure *playing some causal role* in the system's evolution/dynamics; not merely an epiphenomenon. This is the case of the emergent dependence structure. Possibly we need even more than this: Truly *self-organizing emergent structures*. Emergent organizations and orders reproduce, maintain, stabilize themselves through some feedback: either through evolutionary/selective mechanisms or through some form of learning. Otherwise, we do not have a real emergence of causal properties (a new complexity level of organization of the domain).

This is true also for emergence among cognitive/deliberative agents: *the emergent phenomena should feedback on them and reproduce themselves without being understood and deliberated* (Elster 1982). This is the most challenging problem of reconciliation between cognition and emergence: unaware *social functions* impinging on intentional actions (Fig. 3).

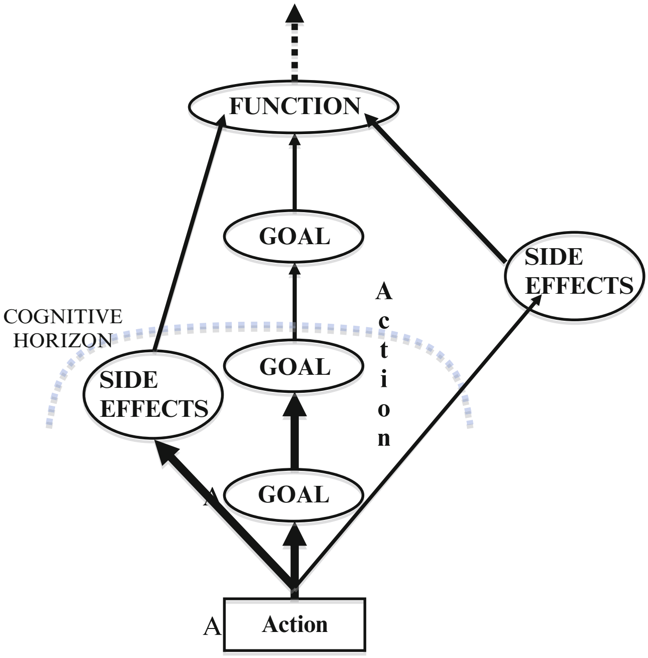


**Fig. 3** Intentions and functions

### *11.1 Reconciling "Emergence" and "Cognition"*

Closing the loop: Emergence and cognition are not incompatible; they are not two alternative approaches to intelligence and cooperation, two competitive paradigms. They must be reconciled:

- first, by regarding cognition itself as a level of emergence: both as an emergence *from sub-symbolic to symbolic* (symbol grounding, emergent symbolic computation), and as a transition *from objective to subjective* representation (awareness) – like in our example of dependence relations – and from *implicit to explicit knowledge*;
- second, recognizing the necessity of going beyond cognition, modeling emergent unaware, functional social phenomena (e.g., unaware cooperation, non-orchestrated problem solving) also among cognitive and planning agents. We have to explain how collective phenomena emerge from individual action and intelligence, and how a collaborative plan can be only partially represented in the mind of the participants, and some part represented in no mind at all.

**Acknowledgements** I would like to thank the members of our research group GOAL http://www.istc.cnr.it/group/goal. I'm in debt with them, which were repeatedly working and discussing with me on these issues. I also want to thank Rainer Reisenzein for his precious revision, comments and criticisms that obliged me to better understand and to make several points more clearly.

## References

Baron-Cohen, S. 1995. *Mindblindness. An essay on autism and theory of mind*. Cambridge, MA: MIT Press.

Batson, C.D. 1991. *The altruism question: Towards a social-psychological answer*. Hillsdale: Lawrence Erlbaum Associates.

Bratman, M.E. 1990. What is intention? In *Intentions in communication*, ed. P.R. Cohen, J. Morgan, and M.E. Pollack. Cambridge, MA: MIT Press.

Bruni, L., and R. Sugden. 2007. The road not taken: How psychology was removed from economics, and how it might be brought back. *Economic Journal* 117(516): 146–173.

Castelfranchi, C. 1992. No more cooperation please! Controversial points about the social structure of verbal interaction. In *AI and cognitive science perspectives on communication*, ed. A. Ortony, J. Slack, and O. Stock. Heidelberg: Springer.

Castelfranchi, C. 1995a. Social commitment: From individual intentions to groups and organizations. In *ICMAS'95 First International conference on multi-agent systems,* 41–49. AAAI-MIT Press.

Castelfranchi, C. 1995b. Guaranties for autonomy in cognitive agent architecture. In *Intelligent agents I*, ed. M.J. Woolridge and N.R. Jennings. Berlin: Springer.

Castelfranchi, C. 1997. Principles of individual social action. In *Contemporary action theory*, ed. R. Tuomela and G. Hintikka. Norwell: Kluwer.

Castelfranchi, C. 2000. Through the agents' minds: Cognitive mediators of social action. *Mind and Society* 2000: 109–140.

Castelfranchi, C. 2001. The theory of social functions. Challenges for multi-agent-based social simulation and multi-agent learning. *Journal of Cognitive Systems Research* 2: 5–38. Elsevier. http://www.cogsci.rpi.edu/~rsun/si-mal/article1.pdf

Castelfranchi, C. 2003a. The micro-macro constitution of power, Protosociology. An International Journal of Interdisciplinary Research. Double Vol. 18–19. In *Understanding the social II – Philosophy of sociality*, ed. Raimo Tuomela, Gerhard Preyer, and Georg Peter.

Castelfranchi, C. 2003b. Grounding we-intentions in individual social attitudes. In *Realism in action – Essays in the philosophy of social sciences*, ed. Matti Sintonen, Petri Ylikoski, and Kaarlo Miller. Dordrecht: Kluwer Publisher.

Castelfranchi, C. 2003c. For a "Cognitive Program": Explicit mental representations for *Homo Oeconomicus*. In *Cognitive processes and economic behavior*, ed. Nicola Dimitri, Marcello Basili, and Itzhak Gilboa. London: Routledge.

Castelfranchi, C. 2011. The "Logic" of power. Hints on how my power becomes his power. *SNAMAS-AISB'11*, 3–9.

Castelfranchi, C. 2012a. Ascribing minds. *Cognitive processing*. Springer.

Castelfranchi, C. 2012b. Goals, the true center of cognition. In *The goals of cognition*, ed. F. Paglieri, L. Tummolini, R. Falcone, and M. Miceli. London: College Publications.

Castelfranchi, C., and R. Conte. 1992. Emergent functionality among intelligent systems: Cooperation within and without minds. *AI & Society* 6: 78–93.

Castelfranchi, C., and R. Falcone. 1997. Delegation conflicts. In *Proceedings of MAAMAW,* ed. M. Boman, and W. van De Welde. Springer-Verlag.

Castelfranchi, C., and F. Paglieri. 2007. The role of beliefs in goal dynamics: Prolegomena to a constructive theory of intentions. *Synthese* 155: 237–263.

Castelfranchi, C., M. Miceli, and A. Cesta. 1992. Dependence relations among autonomous agents. In *Decentralized A.I. – 3*, ed. Y. Demazeau and E. Werner. North Holland: Elsevier.

Conte, R., and C. Castelfranchi. 1995. *Cognitive and social action*. London: UCL Press.

Conte, R., and C. Castelfranchi. 1996. Mind is not enough. Precognitive bases of social interaction. In *Proceedings of the 1992 symposium on simulating societies*, ed. N. Gilbert. London: University College of London Press.

Dennet, D.C. 1981. *Brainstorms*. New York: Harvest Press.

Elster, J. 1982. Marxism, functionalism and game-theory: The case for methodological individualism. *Theory and Society* 11: 453–481.

Falcone, R., and C. Castelfranchi. 1997. "On behalf of ...": Levels of help, levels of delegation and their conflicts, *4th ModelAge workshop*: "Formal Model of Agents", Certosa di Pontignano, Siena.

Frijda, N.H. 1986. *The emotions*. Cambridge: Cambridge University Press.

Garfinkel, H. 1963. A conception of, and experiments with, 'trust' as a condition of stable concerted actions. In *Motivation and social interaction*, ed. O.J. Harvey, 187–238. New York: The Ronald Press.

Grosz, B., and S. Kraus. 1996. Collaborative plans for complex group action. *Artificial Intelligence* 86: 269–357.

Hogg, T., and B.A. Huberman. 1992. *Better than the best: The power of cooperation – Lectures notes in complex systems*. Reading: Addison-Wesley.

Levesque, H.J., P.R. Cohen, and J.H.T. Nunes. 1990. On acting together. In *Proceedings of the 8th national conference on artificial intelligence*, 94–100. San Marco: Kaufmann.

Lewin, K. 1935. *A dynamic theory of personality: Selected papers by Kurt Lewin*. New York/London: McGraw-Hill.

Lorini, E., F. Marzo, and Castelfranchi. 2005. A cognitive model of the altruistic mind. International conference on *Cognitive Economics*, Sofia.

Lorini, E., A. Herzig, and C. Castelfranchi. 2006. Introducing *Attempt* in a modal logic of intentional action. *JELIA* 2006: 280–292.

Lorini, E., N. Troquard, A. Herzig, and C. Castelfranchi. 2007. Delegation and mental states. AAMAS 153.

Mataric, M. 1992. Designing emergent behaviors: From local interactions to collective intelligence. In *Simulation of adaptive behavior*, vol. 2. Cambridge: MIT Press.

McFarland, D. 1983. Intentions as goals, open commentary to Dennet, D.C. Intentional systems in cognitive ethology: the "Panglossian paradigm" defended. *The Behavioural and Brain Sciences* 6: 343–390.

Miller, G., E. Galanter, and K.H. Pribram. 1960. *Plans and the structure of behavior*. New York: Holt, Rinehart & Winston.

Pörn, I. 1989. On the nature of a social order. In *Logic, methodology and philosophy of science*, ed. J.E. Festand et al., 553–567. North-Holland: Elsevier.

Rao, A.S., M.P. Georgeff, and E.A. Sonenberg. 1992. Social plans: A preliminary report. In *Decentralized A. I. 3*, ed. E. Werner and Y. Demazeau. Amsterdam: Elsevier.

Reisenzein, R. 2009. Emotions as metarepresentational states of mind: Naturalizing the belief-desire theory of emotion. *Cognitive Systems Research* 10: 6–20.

Rosenblueth, A., and N. Wiener. 1968. Purposeful and non-purposeful behavior. In *Modern systems research for the behavioral scientist*, ed. W. Buckley. Chicago: Aldine.

Searle, J. 1990. Collective intentions and actions. In *Intentions in communication*, ed. P. Cohen, J. Morgan, and M.E. Pollack. Cambridge, MA: Bradford Books, MIT Press.

Sichman, J.S., R. Conte, C. Castelfranchi, and Y. Demazeau. 1998. A social reasoning mechanism based on dependence networks. In *Readings in agents*, ed. M. Hunhs and M. Singh, 416–421. S. Francisco: Morgan Kaufmann.

Sober, E., and D.S. Wilson. 1998. *Unto others: The evolution and psychology of unselfish behavior*. Cambridge, MA: Harvard University Press.

Steels, L. 1990. Cooperation between distributed agents through self-organization. In *Decentralized AI*, ed. Y. Demazeau and J.P. Mueller. North-Holland: Elsevier.

Tummolini, L. 2010. *Varieties of joint action. A theoretical exploration in the cognitive foundations of social structures.* PhD Thesis, University of Siena.

Tummolini, L., and C. Castelfranchi. 2006. The cognitive and behavioral mediation of institutions: Towards an account of institutional actions. *Cognitive Systems Research* 7(2–3): 307–323.

Tuomela, R. 1993. What is cooperation. *Erkenntnis* 38: 87–101.

Tuomela, R., and K. Miller. 1988. We-intentions. *Philosophical Studies* 53: 115–137.

# Analytical Decomposition of Trust in Terms of Mental and Social Attitudes

**Robert Demolombe**

**Abstract** Trust is defined as a truster's belief in some properties. At the beginning they are to reach a goal and then they are refined in trust in some trustee's property from which the truster can infer that his goal will be reached. This property may be the trustee's ability to bring it about that the goal is reached which can itself be derived from the trustee's intention to reach this goal. Then, we show that this intention may be adopted by the trustee depending on three kinds of social relationships: compliance of norms, mutual commitment with another agent or willingness to act without any compensation.

This analytical decomposition is formalized in a modal logic with a conditional connective. However, the technical details that could prevent an intuitive reading are omitted.

**Keywords** Trust • Ability • Willingness • Compliance • Modal logic

## 1 Introduction

Trust can be analyzed to answer three kinds of questions: "what is the definition of what we call trust?", "on what grounds would someone trust in something?" and "for which purpose trust can be used?". Here we concentrate on trust definition and on some types of trust supports.

There are many definitions of the notion of trust (Castelfranchi and Falcone 2001, 2010; Bacharach and Gambetta 2001; Demolombe 2001, 2004; Demolombe and Liau 2001), nevertheless most of them share the idea that trust is a kind of belief about something. In Jones (2002) and Jones and Firozabadi (2001) Andrew J. I. Jones has shown that these beliefs are a rather complex type of beliefs that combines beliefs in the regularity of some property which may have exception and beliefs in the fact that these exceptions will not arise in the current situation. Since it is not the main topic of this work to characterize the kind of belief which is involved in trust

R. Demolombe (✉)
Institut de Recherche en Informatique de Toulouse, Toulouse, France
e-mail: robert.demolombe@orange.fr

A. Herzig, E. Lorini (eds.), *The Cognitive Foundations of Group Attitudes and Social Interaction*, Studies in the Philosophy of Sociality 5,
DOI 10.1007/978-3-319-21732-1_3

definition we have accepted a very crude definition which is formally represented in epistemic logic by a system of type K (see Chellas 1988).

The supports of belief can be classified into the following categories:

1. series of truster's previous experiences which show the regularity of some property,
2. information transmitted by trusted information sources about this regularity (see Demolombe 2001, 2011),
3. analytical decomposition in function of trust in other properties which are themselves supported by grounds of the type 1, 2 or 3.

The decomposition of type 3 allows to logically derive trust in something from trust in other things. The main topic of this paper is to investigate this decomposition. The formalization help to show what are the trustee's properties that are relevant for this decomposition. Roughly speaking they are mental attitudes or social attitudes of the agents. The formalization in modal logic of these attitudes is only motivated by the objective to propose as far as possible clear definitions of these attitudes and of their relationships. However, we shall not try to give formal detailed definitions of notions like "intention to do" or "attempt to do" which are quite complex and rather controversial. On the contrary we have preferred to leave open these refinements when they have no influence on the decomposition. According to this approach limited information is given about the formal properties of the logic which is presented in the annex.

In the next section, after the informal definitions of the logical framework, we present the starting point of the decomposition and we split it into two categories: trust in the possibility to reach a state of affairs and trust in the possibility to maintain a state of affairs. The decompositions based on these two categories are respectively analyzed in Sects. 3 and 4. In Sect. 5 is presented a comparison with other works and the last section summarizes our conclusions. In the annex some details are given about formal properties.

## 2 Initial Trust Definition

The formal language that will be used in the rest of the paper is defined as follows: *ATOM*: set of atomic propositions, *AGENT*: set of agents, *MODAL*: set of modal operators.

The language is the set of formulas defined by the following BNF:

$$\phi ::= p \mid \neg\phi \mid \phi \vee \phi \mid \phi \Rightarrow \phi \mid M_i\phi$$

where $p$ ranges over *ATOM*, $i$ range over *AGENT* and $M$ ranges over *MODAL*.

The intuitive meaning of the logical connective $\phi \Rightarrow \psi$ is: $\phi$ entails $\psi$.

The set of modal operators *MODAL* and their intuitive meaning is:

$Bel_i\phi$ : $i$ believes that $\phi$ holds.
$Goal_i\phi$ : $i$'s goal is that $\phi$ holds.

$\Box\phi$ : $\phi$ holds now and it will hold always in the future.
$\Diamond\phi$ : there is a future instant where $\phi$ holds (the operator $\Diamond$ is defined from $\Box$ by: $\Diamond\phi \overset{\text{def}}{=} \neg\Box\neg\phi$).
$Attempt_i\phi$ : $i$ attempts to bring it about that $\phi$.[1]
$Int_i\phi$ : $i$'s intention is that $\phi$ holds.
$Obg_j\phi$ : it is obligatory that $j$ brings it about that $\phi$.
$Ask_{i,j}\phi$ : $i$ asks $j$ to bring it about that $\phi$.
$Commit_{j,i}\phi$ : $j$ commits himself with regard to $i$ to bring it about that $\phi$.

In the presented analysis it is assumed that initial trust has practical motivations. That is, trust is a truster's belief in the fact that if he has some given goal, then this goal will be reached. The conditional form of this belief shows that what is trusted is not a property that holds just in the present situation, but rather it holds in every situations where the truster wants to reach this given goal. That is why trust is formally represented by a conditional operator instead of material implication and it has the general form:

$$Bel_i(Goal \Rightarrow GoalReached)$$

where $i$ denotes the truster.

Nevertheless, in most of the real situations this set of situations is restricted to some particular context. For instance, if the truster trusts in the fact that if he wants to take a taxi, then he will find a taxi, his trust may be restricted not to be after midnight and to stay close to downtown. This restriction could be formally represented by the formula:

$$Bel_i(context \Rightarrow (Goal \Rightarrow GoalReached))$$

However, to avoid overly complex formula in the following the context will remain implicit.

This initial trust definition is refined depending on the type of goal. We have considered a first type of goal which is **to reach** a state of affairs. If this state of affairs is denoted by the formula $\phi$, the antecedent of the conditional property is: $\neg\phi \wedge Goal_i\Diamond\phi$, which means that $\phi$ does not hold in the present situation and $i$'s goal is that it holds in the future, and the consequent is: $\Diamond\phi$, which means that the goal $\phi$ will be reached at some future instant. Then, this type of trust is represented by:

(R1) $Bel_i(\neg\phi \wedge Goal_i\Diamond\phi \Rightarrow \Diamond\phi)$

In addition, $i$'s goal is not just that $\phi$ holds at some instant in the future, rather, this instant should happen before a given deadline. For instance, if the truster wants to take a taxi, he is expecting that the taxi will come before some delay. Also, he is aware of the fact that what he trusts may change in a long term future. Then, a more realistic definition would take the form:

[1] The meaning of the operator "to bring it about that $\phi$" can be found in Pörn (1977).

$$Bel_i(Until(d, (\neg\phi \wedge Goal_i(Before(d', \phi)) \Rightarrow Before(d', \phi))))$$

where $Until(d, \psi)$ means that $\psi$ will hold until instant $d$ and $Before(d', \phi)$ means that $\phi$ will hold before the instant $d'$.

However, in the following these temporal refinements will be ignored and we use definition (R1) to avoid overly complex formulas whose intuitive understanding is not easy. The same approach is adopted throughout the rest of the paper.

Examples of (R1).

- $i$ believes that if his car is out of order ($\neg\phi$) and his goal is to have his car repaired ($Goal_i\Diamond\phi$), then it will be repaired ($\Diamond\phi$).
- $i$ believes that if he has a flu and his goal is to be cured, then he will be cured.

Notice that in these examples there is no explicit reference to some trustee.

The second type of goal is **to maintain** a state of affairs $\phi$. Here, the antecedent is denoted by: $\phi \wedge Goal_i\Box\phi$, which means that $\phi$ holds in the present situation and $i$'s goal is that $\phi$ will hold for ever, and the consequent is: $\Box\phi$, which means that $\phi$ will hold for ever. Then, this type of trust is represented by:

(M1) $Bel_i(\phi \wedge Goal_i\Box\phi \Rightarrow \Box\phi)$

Example of (M1). Let's assume that $i$ is visiting a dangerous city. $i$ believes that if he is alive ($\phi$) and his goal is to stay alive ($Goal_i\Box\phi$), then he will stay alive ($\Box\phi$).

In the next sections we analyze from which kinds of trusts (R1) and (M1) can be derived.

## 3 To Reach a State of Affairs

Trust of the type (R1) may be derived from the fact that there exists some agent $j$ such that $i$ believes that if he has the goal to reach a state where $\phi$ holds, then $j$ will attempts to bring it about that $\phi$ (which is represented by (R2)) and $i$ also believes that if $j$ attempts to bring it about that $\phi$, then $\phi$ will hold (represented by (S2)).

(R2) $Bel_i(\neg\phi \wedge Goal_i\Diamond\phi \Rightarrow Attempt_j\phi)$
(S2) $Bel_i(Attempt_j\phi \Rightarrow \Diamond\phi)$

Both (R2) and (S2) are new kinds of trust. The intuitive meaning of (S2) is that $i$ trusts $j$ in his ability to bring it about that $\phi$. If ability is defined as follows:

$$Able_j\phi \overset{\text{def}}{=} Attempt_j\phi \Rightarrow \Diamond\phi$$

(S2) can be represented by: $Bel_i(Able_j\phi)$.

It can be easily shown (see Property RS2 in the Annex) that (R2) and (S2) entail (R1) and this shows that they are a possible analytical decomposition of (R1).

Examples of (R2) and (S2). There exists some agent $j$ such that $i$ believes that if his car is out of order and his goal is to have his car repaired ($\neg\phi \wedge Goal_i\Diamond\phi$), then *j* will attempt to repair his car ($Attempt_j\phi$) and *i* believes that *j* is able to repair his car in the sense that if *j* attempts to repair his car, then his car will be repaired.

Trust of the type (R2) may itself be derived from the fact that *i* believes that if he has the goal to reach the state $\phi$, then *j* will adopt the intention to bring it about that $\phi$ (represented by (R3)) and *i* also believes that if *j* has the intention to bring it about that $\phi$, then *j* will attempt to bring it about that $\phi$. That is represented by (S3)).

(R3) $Bel_i(\neg\phi \wedge Goal_i\Diamond\phi \Rightarrow Int_j\phi)$
(S3) $Bel_i(Int_j\phi \Rightarrow Attempt_j\phi)$

We use the word "determined" to speak about the *j*'s property represented by (S3). This determination property may seem to be obvious, however there are irresolute or indecisive agents who may have the intention to bring it about that $\phi$ and never start to act. This property is formally defined by:

$$Determined_j\phi \overset{\text{def}}{=} Int_j\phi \Rightarrow Attempt_j\phi$$

and trust (S3) is represented by: $Bel_i(Determined_j\phi)$.

It can be shown that (R3) and (S3) entail (R2) (see Property RS23 in the Annex).

Examples of (R3) and (S3). *i* believes that if his car is out of order and his goal is to have his car repaired, then *j* will adopt the intention to repair his car ($Int_j\phi$) and *i* believes that *j* is determined to repair *i*'s car ($Determined_j\phi$) in the sense that if he (*j*) has the intention to repair his car, he will attempt to repair it.

Trust of the type (R3) can be derived from several different kinds of trust. Each one is a possible answer to *i*'s question: what could be a justification of the fact that *j* has adopted the intention to satisfy *i*'s goal?

There are 3 basic answers to this question[2]:

1. *j* is obliged to bring it about that $\phi$
2. if *j* brings it about that $\phi$, then *i* will give to *j* some compensation
3. *j* is willing to help *i* without any compensation

**Case 1.** In case 1 trust (R3) can be derived from the fact that *i* believes that if he has the goal to reach the state $\phi$, then *j* will believe that he (*j*) is obliged to bring it about that $\phi$ (represented by (R4.1)) and *i* also believes that if *j* believes that he is obliged to bring it about that $\phi$, then *j* will adopt the intention to bring it about that $\phi$ (represented by (S4.1)).

(R4.1) $Bel_i(\neg\phi \wedge Goal_i\Diamond\phi \Rightarrow Bel_jObg_j\phi)$
(S4.1) $Bel_i(Bel_jObg_j\phi \Rightarrow Int_j\phi)$

It is worth noting that if *j* ignores that he is obliged to bring it about that $\phi$, there is no chance that this obligation influences *j*'s attitude. That is why in (R4.1) and (S4.1) we have $Bel_jObg_j\phi$ instead of $Obg_j\phi$.

[2] We do not pretend that these three possibilities are exhaustive but we think that they cover most of the situations.

Both (R4.1) and (S4.1) are new kinds of trust. The intuitive meaning of (S4.1) is that $i$ trusts $j$ in his compliance with the obligation to bring it about that $\phi$. If that type of compliance is defined as follows:

$$CompObg_j\phi \overset{\text{def}}{=} Bel_jObg_j\phi \Rightarrow Int_j\phi$$

(S4.1) can be represented by: $Bel_i(CompObg_j\phi)$.

It can be shown that (R4.1) and (S4.1) entail (R3).

Examples of (R4.1) and (S4.1). Here, it is assumed that $j$ is a car mechanic and $i$ is an ambulance driver, and there is a norm which says that if an ambulance is out of order, car mechanics are obliged to repair the ambulance. In this context $i$ believes that if his ambulance is out of order and his goal is to have his ambulance repaired, then $j$ will believe that it is obligatory that he repairs $i$'s ambulance ($Bel_jObg_j\phi$) and $i$ believes that $j$ ordinarily complies with obligation ($CompObg_j\phi$) in the sense that if $j$ believes that it is obligatory that he repairs $i$'s ambulance, then $j$ will adopt the intention to repair it.

Trust (R4.1) can itself be derived from the fact that $i$ believes that if he has the goal to reach the state $\phi$, then there is some agent $k$ who will ask $j$ to bring it about that $\phi$ (represented by (R5.1)) and $i$ also believes that if $k$ asks $j$ to bring it about that $\phi$, then $j$ will believe that he is obliged to bring it about that $\phi$ (represented by (S5.1)).

(R5.1) $Bel_i(\neg\phi \wedge Goal_i\Diamond\phi \Rightarrow Ask_{k,j}\phi)$

(S5.1) $Bel_i(Ask_{k,j}\phi \Rightarrow Bel_jObg_j\phi)$

The intuitive meaning of (S5.1) is that $i$ believes if $k$ order $j$ to bring it about that $\phi$, then $j$ will believe that $k$ has ordered him to bring it about that $\phi$ and $i$ also believes that $k$ has authority (in the sense of "has institutional power" Jones and Sergot 1996) to order $j$ to bring it about that $\phi$. Of course, it is not excluded that $k$ was $i$ himself. If that type of authority is defined as follows:

$$Authorized_{k,j}\phi \overset{\text{def}}{=} Ask_{k,j}\phi \Rightarrow Obg_j\phi$$

(S5.1) can be derived from: $Bel_i(Ask_{k,j}\phi \Rightarrow Bel_j(Ask_{k,j}\phi))$ and $Bel_i(Bel_j(Authorized_{k,j}\phi))$.

It can be shown that (R5.1) and (S5.1) entail (R4.1).

Examples of (R5.1) and (S5.1). Now, it is assumed that there exists a policeman $k$ who has authority to order to the car mechanic $j$ to repair $i$'s ambulance ($Authorized_{k,j}\phi$). In this context $i$ believes that if his ambulance is out of order and his goal is to have his ambulance repaired, then $k$ will ask $j$ to repair it ($Ask_{k,j}\phi$) and $i$ believes that if $k$ asks $j$ to repair $i$'s ambulance, then $j$ will believe that it is obligatory that he repairs it ($Bel_jObg_j\phi$).

**Case 2.** In case 2 (R3) can be derived from the fact that $i$ believes that if he has the goal to reach the state $\phi$, then $j$ will commit with respect to $i$ to bring it about that $\phi$ and $i$ will commit with respect to $j$ to bring it about that $\psi$ (represented by (R4.2)) and $i$ also believes that if this mutual commitment holds, then $j$ will adopt the intention to bring it about that $\phi$ (represented by (S4.2)).

(R4.2) $Bel_i(\neg\phi \wedge Goal_i\Diamond\phi \Rightarrow MutualCommit_{j,i}(\phi, \psi))$
(S4.2) $Bel_i(MutualCommit_{j,i}(\phi, \psi) \Rightarrow Int_j\phi)$

where $MutualCommit_{j,i}(\phi, \psi)$ is defined by:

$$MutualCommit_{j,i}(\phi, \psi) \overset{\text{def}}{=} (Commit_{j,i}\phi) \wedge (Commit_{i,j}\psi)$$

The intuitive meaning of (S4.2) is that if there is a mutual commitment between $j$ and $i$, then $j$ will comply his commitment.

If this compliance is formally defined by:

$$CompCommit_{j,i}(\phi, \psi) \overset{\text{def}}{=} MutualCommit_{j,i}(\phi, \psi) \Rightarrow Int_j\phi$$

(S4.2) can be represented by: $Bel_i(CompCommit_{j,i}(\phi, \psi))$.

It can be shown that (R4.2) and (S4.2) entail (R3).

Examples of (R4.2) and (S4.2). Here, it is no more assumed that $i$ is an ambulance driver. $i$ believes that if his car is out of order and his goal is to have his car repaired, then $j$ will commit himself to repair the car ($Commit_{j,i}\phi$) and $i$ will commit himself to pay $j$ ($Commit_{i,j}\psi$) and $i$ believes that if this mutual commitment between $i$ and $j$ ($MutualCommit_{j,i}(\phi, \psi)$) holds, then $j$ will adopt the intention to repair his car ($Int_j\phi$).

**Case 3**. In case 3 (R3) can be derived from the fact that $i$ believes that if he has the goal to reach the state $\phi$, then $j$ will be aware of his goal (represented by (R4.3)) and $i$ also believes that if $j$ is aware of $i$'s goal, then $j$ will adopt the intention to bring it about that $\phi$ (represented by (S4.3)).

(R4.3) $Bel_i(\neg\phi \wedge Goal_i\Diamond\phi \Rightarrow Bel_jGoal_i\Diamond\phi)$
(S4.3) $Bel_i(Bel_jGoal_i\Diamond\phi \Rightarrow Int_j\phi)$

The intuitive meaning of (S4.3) is that $j$ is willing to satisfy $i$'s goal without any compensation; $j$'s attitude could also be defined as altruist. If $j$'s willingness is defined as follows:

$$Willing_{j,i}\phi \overset{\text{def}}{=} Bel_jGoal_i\Diamond\phi \Rightarrow Int_j\phi$$

(S4.3) can be represented by: $Bel_i(Willing_{j,i}\phi)$.

It can be shown that (R4.3) and (S4.3) entail (R3).

Examples of (R4.3) and (S4.3). Let's assume now that $j$ can observe that $i$'s car is out of order. $i$ believes that if his car is out of order and his goal is to have his car repaired, then $j$ will believe that $i$'s goal is to have his car repaired and $i$ also believes that if $j$ will believe that $i$'s goal is to have his car repaired, then $j$ will adopt the intention to repair his car.

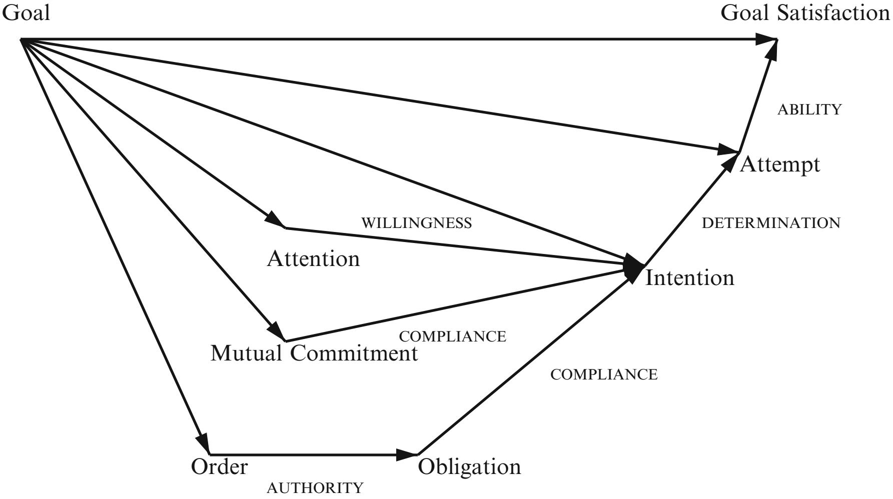


**Fig. 1** Trust analytical decomposition

Figure 1 gives a global picture of the different types of trust and of their relationships.

In the previous analysis of the different types of trust it has implicitly been assumed that the truster $i$ knows who is the trustee $j$ and if $i$ trusts $j$ with respect to several properties, for instance: ability and determination, these properties can be represented by the formulas:

(S2) $Bel_i(Able_j\phi)$
(S3) $Bel_i(Determined_j\phi)$
(R3) $Bel_i(\neg\phi \wedge Goal_i\Diamond\phi \Rightarrow Int_j\phi)$

Due to the logical properties of modality $Bel$, the following can be inferred:

$$Bel_i((Determined_j\phi) \wedge (Able_j\phi) \wedge (\neg\phi \wedge Goal_i\Diamond\phi \Rightarrow Int_j\phi))$$

which entails:

(ExR3) $\exists x Bel_i((Determined_x\phi) \wedge (Able_x\phi) \wedge (\neg\phi \wedge Goal_i\Diamond\phi \Rightarrow Int_x\phi))$

In that formula the variable $x$ which denotes the trustee is interpreted **de re**, that is, the truster $i$ knows who is $x$. However, there may be situations where $i$ believes that there exists some trustee $x$ who holds the properties represented by: $(Determined_x\phi) \wedge (Able_x\phi) \wedge (\neg\phi \wedge Goal_i\Diamond\phi \Rightarrow Int_x\phi)$ even if $i$ does not know such an $x$. In these situations the existential variable $x$ has to be interpreted **de dicto** in the formula:

(ExR'3) $Bel_i\exists x((Determined_x\phi) \wedge (Able_x\phi) \wedge (\neg\phi \wedge Goal_i\Diamond\phi \Rightarrow Int_x\phi))$

It can be shown that both (ExR3) and (ExR'3) entail (R1).

## 4 To Maintain a State of Affairs

If the truster's goal is to maintain a state of affairs we can follow a very similar approach to analyze analytical trust decomposition as in the case where his goal is to reach a state of affairs. However, there are some significant differences.

The decomposition of the initial trust:

(M1) $Bel_i(\phi \wedge Goal_i\Box\phi \Rightarrow \Box\phi)$

in terms of trustee's ability requires two assumptions.

The first one is that no agent who is able to bring it about that $\neg\phi$ will attempt to bring it about that $\neg\phi$ if $i$'s goal is that $\phi$ does not change. The second one (we call it "persistence assumption") is that if the first assumption is satisfied, then $i$ believes that if his goal is to maintain the state of $\phi$, then $\phi$ remain unchanged.

These properties can be formally represented by:

(M2) $\forall x Bel_i(Able_x\neg\phi \rightarrow (\phi \wedge Goal_i\Box\phi \Rightarrow \neg Attempt_x\neg\phi))$
(N2) $\forall x Bel_i(Able_x\neg\phi \rightarrow (\phi \wedge Goal_i\Box\phi \Rightarrow \neg Attempt_x\neg\phi)) \rightarrow$
$Bel_i(\phi \wedge Goal_i\Box\phi \Rightarrow \Box\phi)$

It is obvious that (M2) and (N2) entail (M1).

Notice that $(Able_x\neg\phi) \wedge (Attempt_x\neg\phi)$ entails $\Diamond\neg\phi$ which will mean that $i$'s goal $Goal_i\Box\phi$ fails (see Property MN2 in the Annex). That is why in (M2) it is required that $(\phi \wedge Goal_i\Box\phi \Rightarrow \neg Attempt_x\neg\phi)$ holds.

Another tempting formulation of what is represented by (M2) could be: there is no $x$ such that $i$ believes that $x$ is able to bring it about that $\neg\phi$ and $x$ attempts to bring it about that $\neg\phi$ when $i$'s goal is that $\phi$ does not change:

(M2bis) $\neg\exists x Bel_i(Able_x\neg\phi \wedge (\phi \wedge Goal_i\Box\phi \Rightarrow Attempt_x\neg\phi))$

However, (M2bis) is logically equivalent to:

(M2ter) $\forall x < Bel_i > \neg(Able_x\neg\phi \wedge (\phi \wedge Goal_i\Box\phi \Rightarrow Attempt_x\neg\phi))$

where $< Bel_i >$ is an abbreviation for the possibility operator $\neg Bel_i\neg$, and (M2ter) is consistent with:

(M2qrt) $\forall x < Bel_i > (Able_x\neg\phi \wedge (\phi \wedge Goal_i\Box\phi \Rightarrow Attempt_x\neg\phi))$

which means that it is consistent with what $i$ believes that agents who are able to bring it about that $\neg\phi$ attempt to bring it about that $\neg\phi$ in circumstances where $i$'s goal is that $\phi$ remains unchanged. It is clear that in this situation $i$ cannot trust in the fact that $\phi$ will remain unchanged and that (M2bis) must be rejected.

Another wrong variant of (M2) is:

(M2qnt) $\forall x(Able_x\neg\phi \rightarrow Bel_i(\phi \wedge Goal_i\Box\phi \Rightarrow \neg Attempt_x\neg\phi))$

This formalization is wrong because in (M2qnt) $i$ knows who agents $x$ are but he does not know that these $x$ are able to bring it about that $\neg\phi$. Therefore, $i$ does not know that the set of $x$ who do not attempt to bring it about that $\neg\phi$ contains all the agents who are able to bring it about that $\neg\phi$. That is why (M2qnt) must also be rejected.

In (M2) and (N2) the universally quantified formula $x$ is interpreted *de re*, if it is interpreted *de dicto* we have:

(M'2) $Bel_i \forall x(Able_x \neg\phi \rightarrow (\phi \wedge Goal_i \Box\phi \Rightarrow \neg Attempt_x \neg\phi))$
(N'2) $Bel_i \forall x(Able_x \neg\phi \rightarrow (\phi \wedge Goal_i \Box\phi \Rightarrow \neg Attempt_x \neg\phi)) \rightarrow$
$Bel_i(\phi \wedge Goal_i \Box\phi \Rightarrow \Box\phi)$

Since the two interpretations are very close from a formal point of view, in the following we concentrate only on the *de dicto* interpretation.

Examples of (M'2) and (N'2). In the same context as in the example of (M1), $i$ believes that for every $x$ who is able to kill him ($Able_x \neg\phi$), if $i$'s goal is to stay alive ($\phi \wedge Goal_i \Box\phi$), then $x$ will not attempt to kill him ($\neg Attempt_x \neg\phi$). In addition, $i$ believes that in this situation he will stay alive (N'2). In that example the persistence assumption is quite strong since it excludes situations where $i$ could be killed by accident by someone who is not able to kill him, in the sense that if he attempts to kill $i$ he may kill $i$ but that is not guaranteed.

If the truster $i$ believes that the agents who are determined and able to bring it about that $\neg\phi$ do not adopt the intention to bring it about that $\neg\phi$ when $i$'s goal is that $\phi$ remains unchanged, then $i$ believes that $\phi$ will remain unchanged (see Property MN3 in the Annex). This situation is formally represented by:

(M3) $Bel_i \forall x((Determined_x \neg\phi) \wedge (Able_x \neg\phi) \rightarrow (\phi \wedge Goal_i \Box\phi \Rightarrow \neg Int_x \neg\phi))$
(N3) $Bel_i \forall x((Determined_x \neg\phi) \wedge (Able_x \neg\phi) \rightarrow (\phi \wedge Goal_i \Box\phi \Rightarrow \neg Int_x \neg\phi)) \rightarrow$
$Bel_i(\phi \wedge Goal_i \Box\phi \Rightarrow \Box\phi)$

It is clear that (M3) and (N3) entail (M1).

Examples of (M3) and (N3). The example of (M'2) and (N'2) can be extended here. The only difference is that agents $x$ who are determined to kill $i$ do not adopt the intention to kill him.

It is interesting to observe the formal duality between (M3) and (ExR'3):

(ExR'3) $Bel_i \exists x((Determined_x \phi) \wedge (Able_x \phi) \wedge (\neg\phi \wedge Goal_i \Diamond\phi \Rightarrow Int_x \phi))$

According to this duality, for the decomposition corresponding to **case 1** we have:

(M4.1) $Bel_i \forall x((CompObg_x \neg\phi) \wedge (Determined_x \neg\phi) \wedge (Able_x \neg\phi) \rightarrow$
$(\phi \wedge Goal_i \Box\phi \Rightarrow \neg Bel_x Obg_x \neg\phi))$
(N4.1) $Bel_i \forall x((CompObg_x \neg\phi) \wedge (Determined_x \neg\phi) \wedge (Able_x \neg\phi) \rightarrow$
$(\phi \wedge Goal_i \Box\phi \Rightarrow \neg Bel_x Obg_x \neg\phi)) \rightarrow Bel_i(\phi \wedge Goal_i \Box\phi \Rightarrow \Box\phi)$

Examples of (M4.1) and (N4.1). Like in the examples of (M3) and (N3) it is assumed that there is a criminal organization which can oblige its members to kill somebody. Here, $i$ believes that for every $x$ who complies with the obligations of this organization, if $i$'s goal is to stay alive, then $x$ does not believe that he is obliged to kill $i$.

If the obligations are analyzed as the results of orders given by authorized agents, we have:

(M5.1) $Bel_i \forall x \forall y((Authorized_{y,x} \neg\phi) \wedge (CompObg_x \neg\phi) \wedge (Determined_x \neg\phi) \wedge$
$(Able_x \neg\phi) \rightarrow (\phi \wedge Goal_i \Box\phi \Rightarrow \neg Ask_{y,x} \neg\phi))$

(N5.1) $Bel_i \forall x \forall y((Authorized_{y,x}\neg\phi) \wedge (CompObg_x\neg\phi) \wedge (Determined_x\neg\phi) \wedge (Able_x\neg\phi) \rightarrow (\phi \wedge Goal_i\Box\phi \Rightarrow \neg Ask_{y,x}\neg\phi)) \rightarrow Bel_i(\phi \wedge Goal_i\Box\phi \Rightarrow \Box\phi)$

If $i$ knows who are the authorized agents, instead of (M5.1) which has the form: (M5.1)$Bel_i\forall x\forall y((Authorized_{y,x}\neg\phi)\ldots)$ we have: $\forall y Bel_i\forall x((Authorized_{y,x}\neg\phi)\ldots)$.

Examples of (M5.1) and (N5.1). Like in the previous example, it can be assumed that there are agents $y$ who the authorized agents are in this organization to create the obligation to kill $i$ by asking some $x$ to kill $i$, and these agents do not ask to kill $i$.

For a decomposition corresponding to the **case 2** we have:

(M4.2) $Bel_i\forall x((CompCommit_{x,i}(\neg\phi,\psi) \wedge (Determined_x\neg\phi) \wedge (Able_x\neg\phi) \rightarrow (\phi \wedge Goal_i\Box\phi \Rightarrow \neg MutualCommit_{x,i}(\neg\phi,\psi))$

(N4.2) $Bel_i\forall x((CompCommit_{x,i}(\neg\phi,\psi)\wedge(Determined_x\neg\phi)\wedge(Able_x\neg\phi) \rightarrow (\phi\wedge Goal_i\Box\phi \Rightarrow \neg MutualCommit_{x,i}(\neg\phi,\psi)) \rightarrow Bel_i(\phi \wedge Goal_i\Box\phi \Rightarrow \Box\phi)$

Here $i$'s trust is justified by the fact that there is no mutual commitment between $x$ and $i$ to bring it about that $\neg\phi$.

Examples of (M4.2) and (N4.2). Let's consider now a situation where $i$ is a regular customer of a given hotel. Some days he wants to sleep in the morning ($\phi$) and some other days he wants to be woken up ($\neg\phi$). In this context it may be that if $i$ has a mutual commitment with an employee $x$ of the hotel to be woken up and to give him a tip ($\psi$) in compensation ($MutualCommit_{x,i}(\neg\phi,\psi)$), then $x$ will adopt the intention to wake up $i$ ($CompCommit_{x,i}(\neg\phi,\psi)$). Then, $i$ believes that if $x$ complies with this mutual commitment, if $i$'s goal is not to be woken up, then there will be no such mutual commitment.

For a decomposition corresponding to the **case 3** we have:

(M4.3) $Bel_i\forall x((Willing_{x,i}\neg\phi) \wedge (Determined_x\neg\phi) \wedge (Able_x\neg\phi) \rightarrow (\phi \wedge Goal_i\Box\phi \Rightarrow \neg Bel_x Goal_i\Diamond\neg\phi))$

(N4.3) $Bel_i\forall x((Willing_{x,i}(\neg\phi) \wedge (Determined_x\neg\phi) \wedge (Able_x\neg\phi) \rightarrow (\phi \wedge Goal_i\Box\phi \Rightarrow \neg Bel_x Goal_i\Diamond\neg\phi)) \rightarrow Bel_i(\phi \wedge Goal_i\Box\phi \Rightarrow \Box\phi)$

Here, $i$'s trust is justified by the fact that the agents $x$ who are willing and able to bring it about that $\neg\phi$ do not believe that $i$'s goal is to change the status of $\phi$.

Examples of (M4.3) and (N4.3). In the same example as for (M4.2) and (N4.2), let's assume that instead of agents $x$ who adopt the intention to wake up $i$ in return for a tip we have agents $x$ whose intention to wake up $i$ is only motivated by the fact that they believe that $i$'s goal is to be woken up ($Willing_{x,i}\neg\phi$). In this context, if $i$'s goal is to sleep, $x$ does not believe that his goal is to be woken up ($\neg Bel_x(Goal_i\Diamond\neg\phi)$) and consequently $x$ does not adopt the intention to wake up $i$.

## 5 Comparison with Other Works

At the beginning it was mentioned that we have adopted an extremely crude notion of belief though beliefs play a quite significant role in the notion of trust. In Jones (2002) and Jones and Firozabadi (2001) Andrew J. I. Jones makes the distinction between two kinds of beliefs involved in trust definition: "rule belief" and “conformity belief”. Rule belief expresses a regularity between some state of affairs, formally represented by *context*, and trustee’s behavior (more precisely the fact that the trustee brings it about that $\phi$, represented by $E_j\phi$). This regularity is represented by [3]:

$$Bel_i(context \rightarrow\!\rightarrow E_j\phi)$$

where $\rightarrow\!\rightarrow$ is intended to represent a conditional that tolerates exceptions. Conformity belief expresses that exceptional circumstances will not arise on the occasion concerned.

In Demolombe (2009) Demolombe has proposed a notion of graded trust where a distinction is made between the level of uncertainty $g$ of the truster’s belief, on the one hand, and the regularity level $h$ of the conditional, on the other hand. That is formally represented by:

$$Bel_i^g(\phi \Rightarrow^h \psi)$$

For instance, in the case of trust in determination we could have:

$$Bel_i^g(Int_j\phi \Rightarrow^h Attempt_j\phi)$$

The relationships between the notions of belief presented above deserve further researches.

The idea of trust decomposition has been introduced by Demolombe in Demolombe (2001) for trust in trustee’s epistemic properties. For instance, a “valid” information source is defined as an information source $j$ such that, if $j$ informs the truster $i$ about proposition $\phi$, then $\phi$ holds. This trustee’s property is refined in terms of “sincerity” and “competence”, where agent $j$ is sincere iff if $j$ informs $i$ about $\phi$, then $j$ believes that $\phi$ holds and agent $j$ is “competent” iff if $j$ believes $\phi$, then $\phi$ holds. Then, if $i$ trusts $j$ in his sincerity and his competence, $i$ can infer that he can trust $j$ in his validity. However, in this work there is no reference to $i$’s goal nor to $j$’s intention.

In Castelfranchi and Falcone (2001, 2010) Castelfranchi et al. assume that the truster $i$ has a goal which is to reach a given situation where $\phi$ holds and there exists some other agent $j$, the trustee, such that the truster believes that $j$ can do an action $\alpha$ which has the effect $\phi$ and $j$ has the intention to do this action. This definition is informally characterized by:

- truster’s goal is to reach a situation where the proposition $\phi$ holds
- the action $\alpha$ has the effect that $\phi$ holds
- the trustee has the ability and opportunity to do the action $\alpha$

[3] The notations have been changed in order to make easier the comparison with the presented approach.

- the trustee has the intention to do $\alpha$

A common feature with the presented approach is that trust definition refers to the truster's goal and also to the trustee's ability and intention to reach a state of affairs. However, there are significant differences. The first one is that situations where the truster's goal is to maintain a state of affairs are ignored. The second one is that there is no refinement of an initial definition in terms of other kinds of trust. For instance, there is no attempt to investigate what could justify the fact that the truster has adopted the intention to do action $\alpha$.

This approach has been expressed by Lorini and Demolombe in modal logic in Lorini and Demolombe (2008) with some significant improvements. In particular a notion of obedient agent was introduced which is close to what we have called compliance with obligations and the notion of willingness which is rather close to the kind of willingness we have presented above. It is formally defined by[4]:

$$Will_{j,i}(\alpha) \overset{\text{def}}{=} Goal_j(Bel_j Goal_i Does_{j:\alpha}\top \rightarrow Int_j\alpha)$$

which can be rephrased as: $j$'s goal is that if he believes that $i$'s goal is that $j$ does action $\alpha$, then $j$ adopts the intention to do $\alpha$.

From this notion of willingness is defined "positive trust about willingness" as follows:

$$WTrust(i, j, \alpha, \phi) \overset{\text{def}}{=} Goal_i X\phi \wedge Bel_i(After_{j:\alpha}\phi \wedge Can_j\alpha \wedge Will_{j,i}\alpha)$$

which can be rephrased as: $i$'s goal is that at the next step $\phi$ holds ($X\phi$) and $i$ believes that, after performance of $\alpha$, $\phi$ holds ($After_{j:\alpha}\phi$) and $j$ can do $\alpha$ and $j$ is willing to do $\alpha$.

Here, the difference with our approach is that the only property assigned to the trustee which has a conditional form is his willingness. The other conditions refer to the current situation.

Another difference is that it is implicitly assumed that if $j$ has the intention to do $\alpha$, then he does $\alpha$ and if he does $\alpha$, then $\phi$ will hold. That is, it is implicitly assumed that $j$ is Determined and Able to do $\alpha$ in the sense we have defined. Also, there is no inclusion of the fact that the motivation to adopt the intention to do $\alpha$ may be that there is a mutual commitment between the truster and the trustee.

In this paper is also defined the notion of "negative trust about willingness" which is formally defined by:

$$WTrust(i, j, \neg\alpha, \phi) \overset{\text{def}}{=} Goal_i X\phi \wedge Bel_i(After_{j:\alpha}\neg\phi \wedge Can_j\alpha \wedge Will_{j,i}\neg\alpha)$$

[4] We have simplified the formal definition. In the complete definition there is an additional condition which has been introduced in order to avoid some paradoxes due to material implication.

This type of trust has some common features with trust in maintenance of a state of affairs. The difference is that it “guarantees” that $j$ will not prevent $\phi$ from obtaining, but, it may be that $i$ also believes that another agent than $j$ may prevent $\phi$ from obtaining.

For instance, in the example of the agent who is in a dangerous city and wants to stay alive, the fact the truster believes that agent $j$ will not adopt the intention to kill him “guarantees” that he will not be killed by this agent but that does not “guarantee” that he will not be killed by another agent.

## 6 Conclusion

We have shown how the notion of trust in some property can be grounded on trust in other properties. Truster’s goal to reach a state of affairs or to maintain a state of affairs can be grounded on trust in trustee’s ability to bring it about this state of affairs which can be itself grounded on trustee’s determination to attempt to do what he intends to do. The trustees’s intention may be grounded itself on his compliance of obligations or by mutual interest and commitment with the truster or by willingness to satisfy truster’s goal.

This decomposition has been formalized in conditional logic and in modal logic and we have tried to limit as far as possible the technical details of these logics. In particular, we have adopted a very simple notion of truster’s belief which could be refined in the directions mentioned in the comparison with other works. Another possible improvement could be to go further into the analysis of the temporal dimension, in particular the analysis of how trust changes or persists after the truster has used trust to take decisions and after observation of the effects of these decisions.

**Acknowledgements** I am very grateful to Andrew J.I. Jones for his valuable comments and for his help in the writing of the paper.

## Annex

The axiomatics, in addition to the axiomatics of classical propositional calculus, is defined as follows.

The modal operators $Bel_i$ and $\Box$ obey the axiomatics of a normal modal logic of system K.

For the conditional operator we have the following axiom schemas and inference rules:

(EQUIV) If $\vdash \phi \leftrightarrow \phi'$ and $\vdash \psi \leftrightarrow \psi'$, then $\vdash (\phi \Rightarrow \psi) \rightarrow (\phi' \Rightarrow \psi')$
(TRANS) $(\phi_1 \Rightarrow \phi_2) \wedge (\phi_2 \Rightarrow \phi_3) \rightarrow (\phi_1 \Rightarrow \phi_3)$
(DIST) $(\phi_1 \Rightarrow \phi_2) \rightarrow (\phi_1 \rightarrow \phi_2)$

**Property RS2.**
We have: (R2) $Bel_i(\neg\phi \wedge Goal_i\Diamond\phi \Rightarrow Attempt_j\phi)$ and (S2) $Bel_i(Attempt_j\phi \Rightarrow \Diamond\phi)$ entail (R1) $Bel_i(\neg\phi \wedge Goal_i\Diamond\phi \Rightarrow \Diamond\phi)$.

*Proof.* From the properties of a system K, from (R2) and (S2) we have:

(1) $Bel_i((\neg\phi \wedge Goal_i\Diamond\phi \Rightarrow Attempt_j\phi) \wedge (Attempt_j\phi \Rightarrow \Diamond\phi))$
From (TRANS), we have :
(2) $(\neg\phi \wedge Goal_i\Diamond\phi \Rightarrow Attempt_j\phi) \wedge (Attempt_j\phi \Rightarrow \Diamond\phi) \rightarrow (\neg\phi \wedge Goal_i\Diamond\phi \Rightarrow \Diamond\phi)$
From Necessitation applied to $Bel_i$ and (2) we have:
(3) $Bel_i((\neg\phi \wedge Goal_i\Diamond\phi \Rightarrow Attempt_j\phi) \wedge (Attempt_j\phi \Rightarrow \Diamond\phi) \rightarrow (\neg\phi \wedge Goal_i\Diamond\phi \Rightarrow \Diamond\phi))$
From K and (1) and (3) we have:

(R1) $Bel_i(\neg\phi \wedge Goal_i\Diamond\phi \Rightarrow \Diamond\phi)$

**Property RS23.**
We have: (R3) $Bel_i(\neg\phi \wedge Goal_i\Diamond\phi \Rightarrow Int_j\phi)$,
(S3) $Bel_i(Int_j\phi \Rightarrow Attempt_j\phi)$ and (S2) $Bel_i(Attempt_j\phi \Rightarrow \Diamond\phi)$ entail (R1) $Bel_i(\neg\phi \wedge Goal_i\Diamond\phi \Rightarrow \Diamond\phi)$.

*Proof.* With the same kind of proof as for Property RS2, from (R3) and (S3) we have:

(1) $Bel_i(\neg\phi \wedge Goal_i\Diamond\phi \Rightarrow Attempt_j\phi)$
With the same kind of proof, from (1) and (S2) we have:

(R1) $Bel_i(\neg\phi \wedge Goal_i\Diamond\phi \Rightarrow \Diamond\phi)$.

**Property MN2.**
We have the logical theorem: $Bel_i((Able_x\neg\phi) \wedge (Attempt_x\neg\phi) \rightarrow \Diamond\neg\phi)$.

*Proof.* From (DIST) and *Able* definition we have:

(1) $(Able_x\neg\phi) \rightarrow (Attempt_x\neg\phi) \rightarrow \Diamond\neg\phi)$
Therefore, we have:
(2) $(Able_x\neg\phi) \wedge (Attempt_x\neg\phi) \rightarrow \Diamond\neg\phi)$

Since $Bel_i$ obeys a system K from (2) we have:

$$Bel_i((Able_x\neg\phi) \wedge (Attempt_x\neg\phi) \rightarrow \Diamond\neg\phi).$$

**Property MN3.**
We have the logical theorem: $Bel_i((Determined_x\neg\phi) \wedge (Able_x\neg\phi) \wedge (Int_x\neg\phi) \rightarrow \Diamond\neg\phi)$.

*Proof.* From *Determined* and *Able* definitions, $(Determined_x\neg\phi) \wedge (Able_x\neg\phi)$ is an abbreviation for:

(1) $(Int_x\neg\phi \Rightarrow Attempt_x\neg\phi) \wedge (Attempt_x\neg\phi \Rightarrow \Diamond\neg\phi)$
From (TRANS), (1) entails:

(2) $(\mathit{Determined}_x \neg\phi) \wedge (\mathit{Able}_x \neg\phi) \to (\mathit{Int}_x \neg\phi \Rightarrow \Diamond\neg\phi)$

From (2) and (DIST) we have:

(3) $(\mathit{Determined}_x \neg\phi) \wedge (\mathit{Able}_x \neg\phi) \to (\mathit{Int}_x \neg\phi \to \Diamond\neg\phi)$

And from classical logic (3) entails:

(4) $(\mathit{Determined}_x \neg\phi) \wedge (\mathit{Able}_x \neg\phi) \wedge (\mathit{Int}_x \neg\phi) \to \Diamond\neg\phi$

Since $\mathit{Bel}_i$ obeys a system K, from (4) we have:

(5) $\mathit{Bel}_i((\mathit{Determined}_x \neg\phi) \wedge (\mathit{Able}_x \neg\phi) \wedge (\mathit{Int}_x \neg\phi) \to \Diamond\neg\phi)$

## References

Bacharach, M., and D. Gambetta. 2001. Trust as type detection. In *Trust and deception in virtual societies*, ed. C. Castelfranchi and Y.-H. Tan. Dordrecht/Boston: Kluwer Academic.

Castelfranchi, C., and R. Falcone. 2001. Social trust: A cognitive approach. In *Trust and deception in virtual societies*, ed. C. Castelfranchi and Y.-H. Tan. Dordrecht/Boston: Kluwer Academic.

Castelfranchi, C., and R. Falcone. 2010. *Trust theory: A socio-cognitive and computational model*. Chichester: Wiley.

Chellas, B.F. 1988. *Modal logic: An introduction*. Cambrige: Cambridge University Press.

Demolombe, R. 2001. To trust information sources: A proposal for a modal logical framework. In *Trust and deception in virtual societies*, ed. C. Castelfranchi and Y.-H. Tan. Dordrecht/Boston: Kluwer Academic.

Demolombe, R. 2004. Reasoning about trust: A formal logical framework. In *Trust management: Second international conference iTrust*, LNCS 2995, ed. C. Jensen, S. Poslad, and T. Dimitrakos. Berlin/London: Springer.

Demolombe, R. 2009. Graded trust. In *Proceedings of the trust in agent societies workshop at AAMAS 2009*, Budapest, ed. R. Falcone, S. Barber, J. Sabater-Mir, and M. Singh.

Demolombe, R. 2011. Transitivity and propagation of trust in information sources. An analysis in modal logic. In *Computational logic in multi-agent systems*, LNAI 6814, ed. J. Leite, P. Torroni, T. Agotnes, and L. van der Torre. Berlin/New York: Springer.

Demolombe, R., and C.-J. Liau. 2001. A logic of graded trust and belief fusion. In *Proceedings of 4th workshop on deception, fraud and trust*, Montreal, ed. C. Castelfranci and R. Falcone.

Jones, A.J., and M. Sergot. 1996. A formal characterisation of institutionalised power. *Journal of the Interest Group in Pure and Applied Logics* 4(3): 427–444

Jones, A.J.I. 2002. On the concept of trust. *Decision Support Systems* 33, 225–232.

Jones, A.J.I., and B.S. Firozabadi. 2001. On the characterisation of a trusting agent. Aspects of a formal approach. In *Trust and deception in virtual societies*, ed. C. Castelfranchi and Y.-H. Tan. Dordrecht/Boston: Kluwer Academic.

Lorini, E., and R. Demolombe. 2008. Trust and norms in the context of computer security: A logical formalization. In *Deontic logic in computer science*, LNAI 5076, ed. R. van der Meyden and L. van der Torre. Berlin/New York: Springer.

Pörn, I. 1977. Action theory and social science. Some formal models. *Synthese Library* 120.

# On Modal Logics of Group Belief

**Benoit Gaudou, Andreas Herzig, Dominique Longin, and Emiliano Lorini**

**Abstract** We overview the existing philosophical accounts of group belief, including both aggregative (or reductionist) approaches reducing collective belief to individual beliefs and non-reductionist approaches ascribing beliefs to the group as a whole. We then provide a modal logic of group belief $\mathcal{GL}$ that follows a non-reductionist approach. We compare our group belief logic with the well-known logic of common belief (which is a logic of collective belief in an aggregative sense) and with the logic of group acceptance that has been recently proposed by some of us. Finally, in the spirit of dynamic epistemic logics we propose an extension of $\mathcal{GL}$ by public announcements.

**Keywords** Group belief • Common belief • Modal logic • Epistemic logic

## 1 Introduction

Individual belief has been studied in depth by philosophers and logicians. The latter have developed formal logics, commonly called epistemic logics or doxastic logics, where belief is interpreted as truth in all worlds that the agent considers possible (Hintikka 1962). However, natural language allows not only to ascribe beliefs to individuals, but also to groups. Consider the following examples.

*Example 1 (Tuomela 1992).* The team believes that it will win today's game.

*Example 2 (Gilbert 2002, p. 35).* The United States believe that those responsible for these dreadful acts must be punished.

B. Gaudou (✉) • D. Longin
Institut de recherche en informatique de Toulouse (IRIT), University of Toulouse,
Toulouse, France
e-mail: gaudou@irit.fr

A. Herzig • E. Lorini
University of Toulouse, CNRS-IRIT Universite Paul Sabatier, Toulouse Cedex 9, France

A. Herzig, E. Lorini (eds.), *The Cognitive Foundations of Group Attitudes and Social Interaction*, Studies in the Philosophy of Sociality 5,
DOI 10.1007/978-3-319-21732-1_4

*Example 3 (Meijers 2002, p. 70).* The British believe that the Euro will eventually be introduced in the UK.

Beyond such toy examples, group belief is actually a central concept in multi-agent systems. For example, the reputation value of a seller in an electronic marketplace can be viewed as a belief of a group of agents about that seller (Herzig et al. 2010).

The attribution to a group of an attitude that was previously only studied at the level of individuals is not so obvious and has to be justified. The concept of Intentionality is useful to clarify this point. For Searle, following Brentano (1995) and Husserl, "Intentionality is that property of many mental states and events by which they are directed at or about or of objects and states of affairs in the world" (Searle 1983, p. 1). This characterization can be applied to many mental states such as belief and intention[1]: we believe that the earth is flat, we have the intention to go to the dentist. Thus, belief is an Intentional concept and as such, it is intrinsically ascribed to individuals having a mind and therefore mental representations. From this perspective, a collective or a group not having a mind, it appears fallacious and at best metaphorical to ascribe a belief to a group.

Against this immediate and intuitive idea of individualism, Searle among other authors defends the idea of a genuine collective Intentionality: "the capacity for collective behavior is biologically innate, and the forms of collective Intentionality cannot be eliminated or reduced to something else" (Searle 1995, p. 37). But Searle does not go as far as to defend the idea of a collective mind: the notion of collective Intentionality can be defended without being "committed to the idea that there exists some Hegelian world spirit, a collective consciousness, or something equally implausible" (Searle 1995, p. 25). Tollefsen (2002) gives arguments for collective Intentionality: she points out that groups, organizations, institutions, etc. may be viewed as Intentional agents in the same sense as individual agents. She uses an *interpretationist* approach based on Dennett's notion of *Intentional stance* (Dennett 1987) to defend her point of view: if attitudes like beliefs, intentions, desires, etc. can be ascribed to an agent then this agent is interpretable as an Intentional agent.

In this paper we take for granted the existence of collective Intentionality, and design a modal logic accounting for the ascription of the particular Intentional state of belief to a group. While there are several philosophical accounts of group belief (that we overview in Sect. 2), logical formalizations are much rarer. Two kinds of collective beliefs have been extensively studied in philosophy, artificial intelligence and theoretical computer science: shared belief and common belief (Lewis 1969; Fagin et al. 1995). They are defined in Sect. 2. As we shall see, none of them accounts for the concept of what we are going to call group belief.[2] One

[1] An intention is just a particular attitude having Intentionality. These two notions should not be mixed up and, following Searle, we write Intentionality with a capital 'I' and intention with a small 'i'.

[2] There is also a third kind of collective belief that was studied in artificial intelligence, distributed belief (Fagin et al. 1995), which can be viewed as the belief held by an external observer who knows all the beliefs of the group members. For example, if agent $i$ believes that $\varphi \longrightarrow \psi$ and agent $j$ believes that $\psi \longrightarrow \chi$ then it is distributed belief in the group of agents made up of $i$ and $j$ that $\varphi \longrightarrow \chi$. Such a kind of group belief is therefore in particular implied by individual belief:

of the crucial issues is the relation between individual belief and collective belief: is collective belief determined by individual belief, and if so, how? In particular, does collective belief imply individual belief? Another crucial issue is introspection: does a group belief imply that every member (and more generally every subgroup) is aware of that group belief? And does absence of a group belief imply that every member (and more generally every subgroup) is aware of that absence of group belief? This will be discussed in depth in Sect. 2; for the time being we give an example highlighting these two issues.

*Example 4.* Suppose that agent $i_1$ thinks privately that agent $i_0$ is smart, but that this idea is not widespread. Suppose $i_1$ meets $i_2$ who often claims publicly that agent $i_0$ is dumb; $i_1$ and $i_2$ discuss agent $i_0$ and (for some social reasons) $i_1$ asserts that agent $i_0$ is really a moron, and this point of view is of course shared by $i_2$: after that, a collective belief that $i_0$ is a moron is held by the group made up of $i_1$ and $i_2$. Then agent $i_3$ arrives. Soon the three agents discuss $i_0$. As $i_3$ is the boss of $i_1$ and $i_2$, and as $i_3$ claims that $i_0$ is smart, $i_1$ and $i_2$ quickly agree: the group made up of $i_1$, $i_2$, and $i_3$ holds the collective belief that $i_0$ is smart. The scenario can be continued, yielding an alternating series of group beliefs about $i_0$'s smartness.

Our example illustrates that the collective beliefs held by $i_1$, $i_2$ and $i_3$ as a group contradict the collective beliefs held by $i_1$ and $i_2$, which in turn contradict the individual beliefs of $i_1$ and $i_2$. It also illustrates that individuals and subgroups introspect group beliefs: when $i_1$, $i_2$ and $i_3$ as a group hold the belief that $i_0$ is smart then $i_1$, $i_2$ and $i_3$ are individually aware of that. (This is called positive introspection; in Sect. 2 we shall identify conditions under which group belief also satisfies negative introspection.)

Modal logics for different forms of collective belief exist: there are well-studied logics of shared belief, of common belief and of distributed belief. They however cannot account for our example scenario: first, shared belief of a set of agents that $\varphi$ implies shared belief of every subset that $\varphi$; second, common belief of a set of agents that $\varphi$ implies common belief of every subset that $\varphi$; third, distributed belief is not suitable because it does not satisfy introspection: it might be the case that there is distributed belief that $\varphi$ without any agent believing that there is distributed belief that $\varphi$; our example however requires that a group belief of $i_1$ and $i_2$ that $i_0$ is smart to imply that both $i_1$ and $i_2$ are aware of that group belief.

The preceding considerations motivate the definition of an appropriate logic of group belief, which we undertake in this paper.

The paper is organized as follows. In Sect. 2 we summarize the debate about the ascription of belief to groups. In Sect. 3 we present our logic of group belief. Section 4 extensively discusses the properties of our logic: we assess its properties with respect to the criteria summarized in Sect. 2, compare it to the logic of common belief and to a logic of the (individual and collective) *acceptances* of the members of an institution that has been recently developed (Gaudou et al. 2008; Lorini et al. 2009), and sketch an extension with public announcements. Section 5 concludes.

if $i$ believes that $\varphi$ then it is distributed belief in every group of which $i$ is a member that $\varphi$. This fundamentally different from the concepts that we study here.

## 2 Theories of Collective Belief

Several researchers have tried to reduce collective belief to individual belief. We survey these approaches in Sect. 2.1. But such a reductionist approach cannot capture all aspects of collective belief, and some researchers have therefore followed a non-reductionist approach. We present their theories in Sect. 2.2. Finally, Sect. 2.3 summarizes the main features of a group belief notion.

Throughout the paper we use the following terminology.

- A *collective* is any set of individuals.
- A *group* is a constituted collective: a collective having some structure, some shared goals, rules, etc. An example of a non-constituted collective is the set of all persons having fair hair. See Sect. 2.2.1 for more details on the notion of constituted group.
- An *individual belief* is a belief held by an individual agent. It is therefore private to this agent: no other agent has direct access to this belief.
- A *shared belief* is a belief that is individually held by each agent in a set of agents: "everybody privately believes that $\varphi$". A shared belief is nothing but a conjunction of individual beliefs of the members of this set of agents.
- A *common belief* is the case when every iteration of individual belief holds.[3] So a common belief that $\varphi$ is the case if and only if every agent believes $\varphi$, every agent believes that every agent believes $\varphi$, and so on *ad infinitum*. In the sequel, the term common knowledge refers to a similar definition with knowledge instead of belief.
- A *group belief* is a belief that is held by a constituted collective, alias a group. (Tuomela uses the term *proper group belief* (Tuomela 1992).)
- *Collective belief* is the most general term, subsuming shared belief, common belief and group belief. The collective might be constituted (i.e., it might be a group) or not.

### 2.1 Reductionist Approaches of Collective Belief

Traditionally, collective belief has been viewed as a label of a particular configuration of individual beliefs. This reductionist view of collective belief is called a "summative approach" by Quinton (1976) and Gilbert (1987) and a "statistical" or "aggregative" approach by Tuomela (1992), and is described as an "opinion poll conception" by Meijers (2002). The key point of all these approaches is that such a collective belief is strongly linked to individual beliefs to which it can be reduced, hence the denomination "reductionist approaches". In the sequel we focus on Gilbert's and Tuomela's.

[3] Several authors use the term 'mutual belief' instead of common belief, in particular Tuomela (1992).

### 2.1.1 Gilbert's Account

As a first attempt, Gilbert proposes a simple account that is close to the notion of individual belief and is defined as follows:

**Definition 1 (Simple summative account Gilbert 1987).** A group $J$ collectively believes that $\varphi$ if and only if most of its members believe that $\varphi$.

This account is well-adapted to capture examples such as.

*Example 5 (Tuomela 1992).* Europeans believe that face-to-face discussants should keep at least half a meter apart from each other.

Indeed, if it appears as the result of an opinion poll on Europeans that most of them think (or assert that they think that) face-to-face discussants should keep at least half a meter apart from each other, it is commonly said that Europeans think so.

If we write $|J|$ for the cardinality of the set $J$ and $Bel_i\, \varphi$ for "agent $i$ believes that $\varphi$", then a collective belief of the set of agents $J$ that $\varphi$ is written as the conjunction

$$\bigvee_{J' \subseteq J, |J'| > \frac{|J|}{2}} \bigwedge_{i \in J'} Bel_i\, \varphi$$

where $|J|$ is the cardinality of the set $J$.

Gilbert questions this account by means of the following example.

*Example 6 (Durkheim and Mauss 1963, p. 44).* The Zuni tribe believes that the north is the region of force and destruction.

Suppose that each Zuni believes that the north is the region of force and destruction, and that nobody is aware that his belief is shared (because the Zunis keep their beliefs secret). In this case Definition 1 applies (it is a particular case where $J' = J$). Nevertheless, it seems counterintuitive to say that the Zuni tribe believes that the north is the region of force and destruction. Therefore the opinion poll approach is too weak to capture the notion of group belief.

We can extend the above criticism to take into account links between agents of a group. We can try to propose the following slightly more complex definition: "a group believes $\varphi$ iff every agent of the group believes that $\varphi$ and that every other member believes it too but thinks that they are alone to have this information". This characterization goes one step further by taking into account the set of agents, but the fact that there is still an individual and secret part in this definition prevents it from properly characterizing group belief. For example if every fierce Zuni warrior thinks that $\varphi$ and believes that every other member thinks so, but is not aware that others are aware that he believes this sentence, then no tribe member will dare to assert that there is a group belief about $\varphi$.

To go even further, a complex summative account of collective belief has been proposed by Gilbert, based on the notion of common knowledge, which is a notion that is formally defined, among other works, in philosophy in Lewis (1969, 1972), Schiffer (1972), and Heal (1978), in computer science in Fagin et al. (1995), and in economics in Aumann (1976).

**Definition 2 (Complex summative account (Gilbert 1987)).** A group $J$ collectively believes that $\varphi$ iff:

(1) most of the members of $J$ believe that $\varphi$, and
(2) it is common knowledge in $J$ that (1).

This approach seems better suited to capture the notion of group belief. In particular, a group $J$ with such a belief is aware of its own belief, that is, if $J$ believes $\varphi$ then it is common knowledge in $J$ that $J$ believes $\varphi$. Indeed, if it is common knowledge in $J$ that $\varphi$ then it is common knowledge in $J$ that it is common knowledge in $J$ that $\varphi$. In this sense, an interesting feature of this notion of group belief is its public nature: if $J$ believes $\varphi$ then it is public in $J$ that $J$ believes $\varphi$.

According to that definition, a collective belief of the set of agents $J$ that $\varphi$ has to be written

$$\bigvee_{J' \subseteq J, |J'| > \frac{|J|}{2}} \left( \bigwedge_{i \in J'} Bel_i\, \varphi \wedge CKnow_J \bigwedge_{i \in J'} Bel_i\, \varphi \right)$$

where $Bel_i\, \varphi$ stands for "agent $i$ believes that $\varphi$" and $CKnow_J\, \varphi$ for "it is common knowledge of the agents in $J$ that $\varphi$". If we suppose that knowledge is true then this formula is logically equivalent to $\bigvee_{J' \subseteq J, |J'| > \frac{|J|}{2}} CKnow_J \bigwedge_{i \in J'} Bel_i\, \varphi$.

But this definition is not free from criticism either because it is built from individual beliefs. In particular, it does not allow members of the group to hold individual ('private') beliefs independently from the collective belief: the collective belief imposes that at least the majority actually holds a concordant belief. However, independence from individual beliefs is a particularly interesting feature of a proper notion of group belief, as we are going to explain in Sect. 2.2.

### 2.1.2 Tuomela's We-Belief Account

Tuomela (1992) investigates several collective attitudes (that he calls "we-attitudes"). In particular he proposes an aggregative account of group belief that he calls *shared we-belief*. Situations where such a belief holds are understood as situations where the members of the group may utter "We believe that $\varphi$". His notion of we-belief can be approximately defined as follows.

**Definition 3 (Simple We-belief Account (Tuomela 1992)).** A group $J$ collectively believes $\varphi$ as "We believe that $\varphi$" if and only if every agent $i$ member of $J$ believes

(1) that $\varphi$ and,
(2) that it is commonly believed in $J$ that $\varphi$.

According to Tuomela, this definition suits cases where the set of agents is an aggregate rather than a social, structured group.

*Example 7 (Tuomela 1992).* The Finns believe that sauna originated in Finland.

Each Finn individually believes that Finland is the country of origin of sauna and that this fact is commonly believed by the Finns. When Tuomela considers Finns in the aggregative sense, he understands only the set of agents with the common feature to have Finnish citizenship. No hierarchical link between the Finns is taken into account: when we want to consider Finns as a structured group having institutions and hierarchies between agents, we should use the term Finland instead of Finns, as in the sentence: "Finland declares war against United States".

If we write $Bel_i\, \varphi$ for "agent $i$ believes that $\varphi$" and $CBel_J\, \varphi$ for "it is common belief of $J$ that $\varphi$" then the fact that group $J$ collectively believes that $\varphi$ as "We believe that $\varphi$" (i.e., that members of $J$ share a we-belief that $\varphi$) is formally written as:

$$\bigwedge_{i \in J} Bel_i\ (\varphi \wedge CBel_J\, \varphi)$$

We note in passing that if we suppose that the logic of individual belief and common belief is the standard normal modal logic $\text{KD45}_n^C$ then this formula is logically equivalent to $CBel_J\, \varphi$.

### 2.1.3 On the Insufficiencies of Reductionist Approaches

Although the above approaches seem to suffice to represent many cases of collective belief, they do not account for all of them. In this section we present the main arguments provided by, among others, Gilbert and Meijers highlighting several insufficiencies.

A first argument against both simple and complex summative accounts (cf. Definitions 1 and 2) can be illustrated by the following example.

*Example 8 (Gilbert 1987).* It is probably common knowledge in the population of adults who have red hair and are over six feet tall that most of them believe that fire burns.

It seems too strong to ascribe a group belief that fire burns to this set of human beings. Indeed, the population of adults who have red hair and are over six feet tall does not necessarily constitute a *group*: there are no social relationships between the members of this large population, and there is no common identity.

Second, group belief should have a binding effect on the group members. Consider the following example.

*Example 9 (Tuomela 1992).* The Government believes that war against Iraq will begin soon.

Here, every agent *qua* member of the government has to express and act according to the fact that the government believes that war in Iraq is imminent: each member should defend and argue for this belief if it is challenged by another agent, as if it was her personal and private belief. This should however not imply anything with respect to her private beliefs. Moreover, every agent takes it also for granted that every other government member will act so and thus cannot change her mind at the social level (taken as the mental attitudes that he expresses) without any discussion with other government members: every agent is bound to this group belief, and changing her mind at the social level should be the result of a group consensus.

The previous summative accounts do not have anything to say about that binding nature: indeed, nothing in common belief as defined above has a binding or persistent feature. Moreover as soon as an agent privately changes her mind[4], for any reason or evidence and independently of other group members, the common belief vanishes.

Thirdly, Meijers (2002) argues that this commitment to group belief is conditioned by its acceptance by other members and their commitment to it. Consider the following example that is inspired by the Prisoner's Dilemma.

*Example 10.* Two criminals were arrested. Both publicly claim that they are innocent. We can thus ascribe to them a group belief that they are innocent. They are examined separately: they still claim that they are innocent and that their partner is also innocent. But if a policeman informs one of them that the other has defected (and if that prisoner believes the policeman), then the latter will typically consider that their binding commitment is broken and might defect, too.

As our example illustrates, every group member is committed to defend the group belief in front of anyone, but as soon as one member violates this commitment, other members no longer have to defend the group belief since the constituted group does not exist anymore. Just as the above feature, this aspect cannot be understood on the basis of a reductionist account of collective belief.

Fourth, the last and perhaps most important argument against the reductionist approaches Meijers (2002) (and also cited by Gilbert (1987) and Tuomela (1992)) is that group belief should be independent from individual beliefs. Tuomela gives the following example.

*Example 11 (Tuomela 1992).* The Communist Party of Ruritania believes that capitalist countries will soon perish (but none of its members really believes so).

This example highlights that a group can believe a statement without any member believing it privately. Of course, this is an extreme case (called 'spurious collective

[4] In Gilbert's simple account, if only one agent changes her mind then the common belief vanishes. In complex accounts, a change of mind of several agents (at worst most of the agents) is needed. However, it is still the case that some agents privately making up their minds may modify the collective belief, independently of any discussion and consensus.

belief' by Tuomela (1992)). However, a correct account of group belief should leave room for situations such as the above one, and a group belief that $\varphi$ should not systematically imply individual beliefs of the group members that $\varphi$. Conversely, Example 6 shows that although every member of the Zuni tribe believes that the north is the region of force and destruction, we cannot ascribe a group belief to the tribe because its members keep their feelings and beliefs secret. Thus, a conjunction of individual beliefs does not necessarily imply group beliefs either. To sum up, a comprehensive account of group belief should keep these two notions independent, in the sense that any of their Boolean combinations should be consistent.

This desideratum was already expressed by Durkheim (1982), who asserts that any proper group belief must be "external to individual consciousness". Indeed, a group belief is often the result of a negotiation, a deliberation or a persuasion process and thus of a consensus between two or more parties with very different viewpoints. It can even be the result of more or less ethical processes such as propaganda or threat (as in Example 11). Durkheim's criterion allows one to handle cases where collective belief is the result of a discussion and where a compromise between each disputant has been reached, as in the following example.

*Example 12 (Meijers 2003).* A selection committee can believe that a particular candidate is the best candidate for the job, without any of its members believing this individually. Each of them could have a different candidate as their first choice. However, in their role of members of the committee they believe the selected candidate to be the most appropriate for the job.

In this example, the group belief that a particular candidate is the best choice typically results from a voting procedure. Such group belief generating procedures are studied in the field of social choice theory Taylor (2005). We leave aside these more elaborate group belief formation mechanisms in the present logical analysis of group belief and only adopt a very simple principle in the logic of Sect. 3: unanimity.

## 2.2 Non-reductionist Accounts: Towards Group Belief

The above criticisms were the starting point for a new approach to group belief that was mainly led by Gilbert (1987), who considers group belief as a primitive concept that cannot be reduced to individual attitudes. Such accounts are called *non-reductionist* in the sequel.

### 2.2.1 The Plural Subject Account

Let us begin by an example.

*Example 13 (The poetry group (Gilbert 1987)).* A group of people meet regularly at one member's house to discuss poetry. The format followed when they meet, which evolved informally over time, is as follows. A poem by a contemporary poet is read out. Each participant feels free to make suggestions about how to interpret and evaluate the poem. Others respond, as they see fit, to the suggestions that are made. An opposing view might be put forward, or data adduced to support or refute a suggestion which has been made.

From this discussion a consensual view of the poem will emerge. It will represent the view of the group or the collective opinion about this poem, i.e., it is the belief of the group about this poem. Although this attitude appears to be a belief, it does not have the same properties as collective belief in the summative sense.

In opposition to the summative approach, Gilbert proposes in Gilbert (1989) the following characterization of what she calls group belief.[5]

**Definition 4 (The plural subject account (Gilbert 1989)).**

(1) *A group $J$ believes that* $\varphi$ if and only if the members of $J$ jointly accept that $\varphi$.
(2) The members of $J$ *jointly accept* that $\varphi$ if and only if it is common knowledge in $J$ that the members of $J$ individually have intentionally and openly expressed their willingness jointly to accept that $\varphi$ with the other members of $J$.

A first thing to observe is that the concept of individual belief is absent from this definition. There is an important reason for untying group beliefs from individual attitudes (and individual beliefs in particular) and for not reducing the former to the latter. While group belief is *public* with respect to the members of the group, individual attitudes of agents are *private*, i.e. inaccessible to other agents. We can only have access to them in an indirect way, by observing agents' behaviors and actions, and by trying to interpret them.

Another important aspect of Gilbert's notion of group belief is that it entails both an identification and a mutual recognition with respect to the same group. That is, a group belief of the agents in group $J$ is based on the fact that the agents in $J$ identify themselves as members of this group, recognize each other as members of this group, and accept certain things to stand as the view of the group. We call this kind of set of agents a *constituted group*. This aspect is explicitly stated in several parts of Gilbert's book. For instance, just before proposing her 'official' notion of group belief she says (Gilbert 1989, p. 289) "I suggest that what is both logically necessary and logically sufficient for the truth of the ascription of group belief here is, roughly, that all or most members of the group have expressed willingness to let a certain view to stand as the view of the group." Some pages later she provides additional clarification: "There are, evidently, various ways of describing the situation of those who have jointly accepted a view with certain others. We may say they have undertaken to express a certain view when acting within, or as a representative of the group. One might even say that someone has accepted a view *qua member of a certain group*." (Gilbert 1989, p. 304)

[5] We note that Gilbert uses indistinctly the terms "collective belief" and "group belief" (Gilbert 1989).

It is also to be noted that Gilbert's notion of group belief implies a common belief of the group about the existence of this group belief: as joint acceptance requires common knowledge of every agent about her willingness to accept the proposition, we can deduce that every member is aware of the group belief, and even that there is common knowledge of this, which also implies common belief of the existence of the group belief. So common knowledge is only about the group belief itself. In contrast, the above Definition 2 requires common knowledge on individual beliefs that $\varphi$ for the summative collective belief that $\varphi$.

Finally, it is to be noted that in Gilbert's mind, the term acceptance has to be taken in the commonsensical use and not in the philosophical sense, where acceptance is opposed to belief.[6] Moreover, Tuomela (1992) remarks that his own definition is circular: the word *joint acceptance* is used in its own definition. In reaction to this and the above remark, Gilbert proposed a slightly different characterization of group belief.

**Definition 5 (Gilbert 2002).** The members of a population *P* collectively believe that $\varphi$ if and only if they are *jointly committed* to believe that $\varphi$ as a body.

Joint commitment is a persistent positive attitude toward a decision taken by a group, similarly to personal commitment to stick to an intention until it is fulfilled (Bratman 1987; Cohen and Levesque 1990). As far as Gilbert is concerned, a joint commitment is formed by the expression by every group member of his readiness to be committed to the view of the group. We note that there is common knowledge in the group of this joint commitment; see Part III of Gilbert (1996) for more details.[7]

Tuomela (1992) generalizes the previous approach by introducing the concept of *operative members*. These are particular members of a group that can impose their beliefs to other members of the group. We do not present his approach in detail here because his concept of operative member does not seem to be a necessary ingredient of the concept of group belief: in many cases the set of operative members of a group is the whole group, e.g. in the case of inhabitants of a country, or families with children old enough to take part in the decision making process (Tuomela 1992).

### 2.2.2 Against Gilbert's Plural Subject Account: Belief *vs* Acceptance

Some authors reject that group beliefs in Gilbert's sense are really beliefs, and claim that what Gilbert calls belief is another kind of attitude, namely acceptance. Among those who are named *rejectionists* by Gilbert (2002), we can cite K. Brad Wray (2001, 2003), Anthonie Meijers (1999, 2002, 2003) and Raimo Tuomela (2000).

[6] She has explicitly made a distinction between her notion of group belief and acceptance in the philosophical sense in Gilbert (2002). Nevertheless, some authors think that Gilbert's group belief is acceptance in the philosophical sense (see the next section for more details).

[7] We note that this imposes some hypotheses on communication: the channel is perfect, agents are always aware of every information they got, and there is no misunderstanding.

Most of them neither reject the notion of collective Intentionality nor the idea of ascribing mental attitudes to a group, nor are they opposed to Gilbert's plural subject account (our Definition 4). It is rather the nature of the phenomenon the plural subject account describes that is controversial. Everybody agrees that the group belief resulting from the plural subject account is a collective doxastic state, but while Gilbert argues that it corresponds to a form of belief, K. Brad Wray responds (Wray 2001) that "(...) the phenomenon that concerns Gilbert is a species of acceptance [rather than belief]". Thus rejectionnists agree with non-reductionist approaches of group belief or more generally of collective Intentionality, but they consider that Gilbert's group belief is not a kind of group belief but rather a kind of group acceptance. It is outside the scope of the present paper to go into the details of this debate, and we refer interested readers to Cohen (1989) and Hakli (2006) or to our overview in Lorini et al. (2009). We will however compare in Sect. 4.3 our logic of group belief with the logic of acceptance that we have proposed in Lorini et al. (2009).

## *2.3 Toward a Formal Characterization of Group Belief*

In the next section we will propose a formalization of group belief in terms of a logic having a modal operator of group belief. Before this we are going to sum up the main points of the discussion of the present section: they are going to guide our formalization. The five criteria are mainly extracted from Gilbert's plural subject account of group belief, and we view them as proper features of group belief. They also allow to distinguish genuine group belief from aggregate collective belief (that we are going to identify in the sequel with common belief).

**Group belief has a binding force.** As emphasized by Gilbert (1989), a group belief held by a set of agents $J$ entails that the agents in $J$ identify and recognize themselves as members of the same group, and accept certain things to stand as the view of the group. This might be called the *binding force* of group belief: if the agents in $J$ hold a certain group belief then they think of themselves as members of the same group, and they are bound by a common identity. From this perspective, when the agents in $J$ hold a group belief then $J$ should not be simply conceived as a set of agents. This binding force is completely ignored in the summative account that we have discussed in Sect. 2.1. It appears explicitly in our official reading of the modal operator of group belief to be introduced in the next section.

**No combination of individual beliefs implies group belief.** In our view, this is the major argument against the summative approaches. It is inspired by Durkheim (1982), and is one of the most important contributions of Gilbert's account. This property is typically illustrated by Example 6 about the Zuni tribe: the fact that every member of a set of agents believes that $\varphi$ is true is not a sufficient condition for the group belief that $\varphi$.

**Group belief does not imply individual belief.** Conversely, this property says that every member of a set $J$ of agents believing that $\varphi$ is not a necessary condition for the group belief that $\varphi$. Together with the previous property, this property means that there should be no entailment link between the group belief operator and the individual belief operator. Thus, our group belief operator will be able to account for Tuomela's "spurious collective beliefs" (cf. Example 11). This contrasts with common belief (on which are based Gilbert's complex summative account and Tuomela's account), which implies individual, private beliefs.

**There is a commitment to group belief.** As soon as a group belief has been established, even if some group members disagree with this belief, they must act in compliance with it, i.e., they are committed to this belief. When they violate this commitment they are liable for sanctions, ranging from blames (Gilbert 1987) to exclusion from the group (Tuomela 1992). In the sequel, we will not consider a fine sanction system; instead we logically forbid violation of a group belief by a member of a constituted group, in the sense that if a set of agents publicly holds beliefs that are jointly inconsistent then it cannot constitute a group.

**The group members hold a common belief about group beliefs.** One of the major criticisms against the "simple summative approach" is that every group member may individually believe that $\varphi$ without any collective belief that $\varphi$ because agents are not aware of the other agents' beliefs. A kind of common belief is thus necessary, but not about the content of the group belief (as in the "complex summative approach"), but rather about the group belief itself: the group belief is public for every member of this group. Tuomela (1992) defends this thesis arguing that group belief is built due to a joint and intentional group action: if such a group action has occurred then this is known by each agent (and is even common knowledge). This entails that every member of the group is aware of all group beliefs.

The above five previous requirements constitute the main features of a group belief operator. They are precious guidelines in our logical formalization of group belief. Nevertheless they will not be translated directly into the logical axioms to be presented in Sect. 3.3: the positive requirements will be derived from the logical principles that we will adopt, and the negative requirements will not be derivable due to some principles that we are going to reject. This will be discussed in Sect. 4.2.1.

## 3 A Logic of Group Belief

We now turn to a logical formalization of Gilbert's non-reductionist group belief. Our logic is based on the standard multi-modal logic of individual belief KD45$_n$ (Fagin et al. 1995). We augment this logic by a modal operator of group belief and call the resulting logic the *logic of group belief*, noted $\mathcal{GL}$. We first present the Kripke semantics of our logic and then axiomatize its validities. In Sect. 4 we will extend $\mathcal{GL}$ with operators of common belief and of public announcement.

## 3.1 Syntax

Let $AGT$ be a finite set of agents. We use $i, j, \ldots$ to denote elements of $AGT$, and $J, J', \ldots$ to denote non-empty subsets of $AGT$.[8] Let $ATM = \{p, q, \ldots\}$ be a countable set of propositional letters. The language of our logic $\mathcal{GL}$ is defined by the following grammar:

$$\varphi ::= p \mid \neg\varphi \mid \varphi \vee \varphi \mid GBel_J\, \varphi$$

where $p$ ranges over $ATM$ and $J$ over $2^{AGT} \setminus \{\emptyset\}$. When $J$ is a singleton then we write $GBel_i\, \varphi$ instead of $GBel_{\{i\}}\, \varphi$. We identify $GBel_i\, \varphi$ with the individual belief $Bel_i\, \varphi$. The Boolean connectives $\wedge$, $\rightarrow$, $\leftrightarrow$, $\top$ and $\bot$ are defined from $\vee$ and $\neg$ in the usual manner.

$GBel_J\, \varphi$ reads "while the set of agents $J$ is a constituted group then $J$ believes as a whole that $\varphi$ holds". Therefore $GBel_J\, \bot$ may be read "The set of agents $J$ is not a constituted group" and $\neg GBel_J\, \bot \wedge GBel_J\, \varphi$ may be read "the set of agents $J$ is a constituted group and believes $\varphi$", or "the group $J$ believes $\varphi$". In the case of a singleton $\{i\}$, we identify the group belief operator $GBel_{\{i\}}$ with the individual belief operator, avoiding thus a particular operator of individual belief. $GBel_{\{i\}}\, \varphi$ reads "agent $i$ believes that $\varphi$ holds". We sometimes also say that $i$ individually believes that $\varphi$, or that $i$ privately believes that $\varphi$. For convenience, we write $GBel_i\, \varphi$ for $GBel_{\{i\}}\, \varphi$.

## 3.2 Semantics

A model $M$ of the logic of group belief includes a nonempty set of possible worlds $W$ and a valuation function $\mathscr{V} : ATM \longrightarrow 2^W$ associating to each propositional letter $p$ the set of worlds where $p$ is true. Models moreover contain accessibility relations that will be detailed in the sequel.

To each possible world $w$ and each non-empty set of agents $J \subseteq AGT$, we associate the set of possible worlds $\mathscr{G}_J(w)$ that are consistent with all propositions believed in world $w$ by $J$. $\mathscr{G}_J(w)$ contains those worlds that are consistent with what is believed by $J$. Formally, we have a mapping $\mathscr{G} : (2^{AGT} \setminus \{\emptyset\}) \longrightarrow 2^{W \times W}$ associating an accessibility relation to each non-empty subset of $AGT$. For convenience, we write $\mathscr{G}_i$ instead of $\mathscr{G}_{\{i\}}$.

For a given model $M$, the truth condition for $GBel_J$ stipulates that $\varphi$ is believed by $J$ at $w$, noted $M, w \models GBel_J\, \varphi$, if and only if $\varphi$ holds in every world that $J$ can access from $w$ via $\mathscr{G}_J$:

$$M, w \models GBel_J\, \varphi \text{ iff } M, w' \models \varphi \text{ for every } w' \in \mathscr{G}_J(w).$$

[8] It does not make too much sense to talk about the belief of an empty set of agents. We refer to Ågotnes (2012) for a thorough investigation of the formal issues related to empty groups.

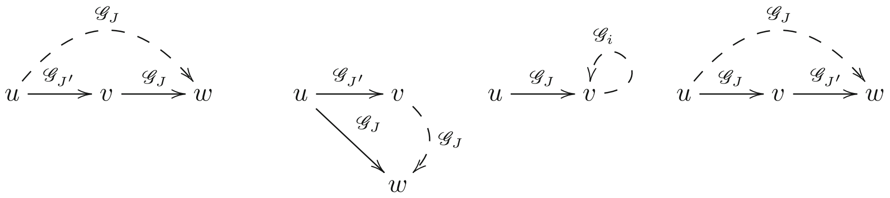


**Fig. 1** Schemas for conditions 1, 2, 3 and 4

We impose the following constraints on accessibility relations of $\mathcal{GL}$ models, for sets of agents $J$ and $J'$ such that $J' \subseteq J$ and $J' \neq \emptyset$:

1. if $u \mathscr{G}_{J'} v$ and $v \mathscr{G}_J w$ then $u \mathscr{G}_J w$;
2. if $u \mathscr{G}_{J'} v$ and $u \mathscr{G}_J w$ then $v \mathscr{G}_J w$;
3. if $u \mathscr{G}_J v$ then there is $i \in J$ such that $v \mathscr{G}_i v$;
4. if $u \mathscr{G}_J v$ and $v \mathscr{G}_{J'} w$ then $u \mathscr{G}_J w$.

Constraint 1 stipulates that agents of a subset $J'$ of the set $J$ are aware of what is collectively believed by the group $J$: whenever $w$ is a world for which it is believed by $J'$ that all $J$-believed propositions hold in $w$, then all $J$-believed propositions indeed hold in $w$ (Fig. 1). Similarly, 2 expresses that subgroups are aware of what is not believed in the group, too (Fig. 1). Together, these two constraints are a kind of *attention* property: each subgroup is aware of what is believed (and not believed) by the group. This is justified by Gilbert's hypothesis that the commitment towards a group belief is common knowledge. 1 and 2 can be put together: if $u\mathscr{G}_{J'}v$ then $\mathscr{G}_J(u) = \mathscr{G}_J(v)$, i.e., if $u\mathscr{G}_{J'}v$ then what is believed by $J$ at $u$ is the same as what is believed by $J$ at $v$. From 1 and 2 it also follows that $\mathscr{G}_J$ is transitive and Euclidian. 3 says that it is believed by a group $J$ that if a proposition is believed by each of $J$'s members then it is true, too. This is a kind of unanimous adoption of group belief. 4 says that if a formula is believed by a group $J$ then it is believed by $J$ that this information is believed by every subgroup of $J$ (Fig. 1).

The accessibility relations $\mathscr{G}_J$ are not necessarily serial: seriality of $\mathscr{G}_J$ means that $J$ is a constituted group. However, we assume that individual agents are rational and thus that their beliefs stay consistent. We therefore impose:

5. $\mathscr{G}_i$ is serial, for every $i \in AGT$.

In words, at least one world exists that is consistent with the set of individually believed propositions.

*Remark 1.* Note that our constraints do not guarantee that $\mathscr{G}_{J'} \subseteq \mathscr{G}_J$, for $J' \subseteq J$. They do not guarantee either that when $u\mathscr{G}_{J'}v$ and $u\mathscr{G}_J w$ then $v\mathscr{G}_{J'}w$: together with 1, it would imply that $\mathscr{G}_{J'}(u) \subseteq \mathscr{G}_J(u)$ as soon as $\mathscr{G}_J(u) \neq \emptyset$.

**Definition 6.**

- $\varphi$ is true in $M$ ($M \models \varphi$) iff $M, w \models \varphi$ for every $w \in W$.
- $\varphi$ is valid in a class of models $\mathsf{C}$ (noted $\models_{\mathsf{C}} \varphi$) iff $M \models \varphi$ for every $M \in \mathsf{C}$.
- $S \models_{\mathsf{C}} \varphi$ iff for every $M \in \mathsf{C}$, if $M \models \psi$ for every $\psi \in S$ then $M \models \varphi$.

The class of all Kripke models satisfying 1, 2, 3, 4 and 5 is called $\mathcal{GL}$. We write $\models_{\mathcal{GL}} \varphi$ when a formula $\varphi$ is valid in the class $\mathcal{GL}$.

## 3.3 Axiomatics

The validities of $\mathcal{GL}$ are axiomatized as follows:

$$\frac{\varphi}{GBel_J\, \varphi} \tag{RN$_{GBel_J}$}$$

$$GBel_J\, (\varphi \to \psi) \to (GBel_J\, \varphi \to GBel_J\, \psi) \tag{K$_{GBel_J}$}$$

$$GBel_J\, \varphi \to GBel_{J'}\, GBel_J\, \varphi \quad \text{if} J' \subseteq J \tag{IN+}$$

$$\neg GBel_J\, \varphi \to GBel_{J'}\, \neg GBel_J\, \varphi \quad \text{if} J' \subseteq J \tag{IN−}$$

$$GBel_J \left( \left( \bigwedge_{i \in J} GBel_i\, \varphi \right) \to \varphi \right) \tag{UNA}$$

$$GBel_J\, \varphi \to GBel_J\, GBel_{J'}\, \varphi \quad \text{if} J' \subseteq J \tag{AGR}$$

$$GBel_i\, \varphi \to \neg GBel_i\, \neg\varphi \tag{D$_{GBel_i}$}$$

The last five axioms respectively correspond to constraints 1–5.

The axioms of *positive* and *negative introspection* (IN+) and (IN−) correspond to constraints 1 and 2, and express that if a proposition $\varphi$ is believed (resp. not believed) by the set of agents $J$ then it is believed by each subset that $\varphi$ is believed (resp. not believed) by $J$. This is due to the public character of the group belief operator. In particular, each agent $i$ member of $J$ is aware of what is believed (resp. not believed) by the group $J$: if $i \in J$ then both $GBel_J\, \varphi \to GBel_i\, GBel_J\, \varphi$ and $\neg GBel_J\, \varphi \to GBel_i\, \neg GBel_J\, \varphi$ (set $J = \{i\}$ to see this). The schemas (IN+) and (IN−) therefore generalize the positive and negative introspection axioms for individual belief.

The axiom (UNA) corresponds to the semantic constraint 3, and expresses that it is collectively believed by $J$ that if every member of $J$ individually believes $\varphi$ then $\varphi$ is true. It is important to remark here that the formula $GBel_J\, GBel_i\, \varphi$ (with $i \in J$) has a particular status. We consider that this formula has as primary origin the expression of $i$'s belief that $\varphi$ holds. Following speech act theory (Searle 1969), the assertion of $\varphi$ by $i$ counts as the public expression of his belief that $\varphi$. We observe that other group members do not have any access to the truth of $GBel_i\, \varphi$. Therefore this individual belief expressed to the group is automatically believed by the group. (This is a shortcut, because it presupposes a perfect communication channel.) In the case where the acceptance of this fact induces an inconsistency with previous group beliefs we consider that the set of agents is no longer a constituted group. We stress that $i$'s public expression of his belief that $\varphi$ in front of group $J$ neither implies that $i$ privately believes that $\varphi$, nor that the members of $J$ privately believe that $\varphi$.

The axiom (AGR) corresponds to the semantic constraint 4, and is an agreement axiom: it says that if $\varphi$ is believed by $J$ then it is believed by $J$ that the formula is believed by each subset $J'$ of $J$. Note that this does not imply that $\varphi$ is actually believed by every subset $J'$ of $J$, i.e., (AGR) does not entail $GBel_J\, \varphi \to GBel_{J'}\, \varphi$. In particular, the fact that $\varphi$ is believed by $J$ does not imply that the members of $J$ individually believe that $\varphi$, i.e., $GBel_J\, \varphi \to GBel_i\, \varphi$ is invalid (regardless whether $i \in J$ or not). Thanks to this axiom we have the following theorems:

$$GBel_J\, (\varphi \wedge GBel_i\, \neg\varphi) \to GBel_J\, \bot \quad \text{if } i \in J \tag{1}$$

$$GBel_J\, (GBel_i\, \varphi \wedge GBel_j\, GBel_i\, \neg\varphi) \to GBel_J\, \bot \quad \text{if } i, j \in J \tag{2}$$

The former theorem highlights a property of Gilbert's notion of group belief (or acceptance) according to which it is implausible that a constituted group $J$ agrees to have a collective view that $\varphi$ and, at the same time, agrees that someone in $J$ has a dissident point of view, i.e., someone in $J$ believes that $\neg\varphi$.

Together, (AGR) and (UNA) entail the following theorem:

$$GBel_J\, \varphi \leftrightarrow (GBel_J \bigwedge_{i \in J} GBel_i\, \varphi) \tag{3}$$

Thus it is believed by a group of agents $J$ that $\varphi$ holds if and only if it is believed by $J$ that each of its members believes $\varphi$. This illustrates the process of group belief establishment by consensus that is 'built in' in our logic.

Note that the conjunction $\neg GBel_i\, \varphi \wedge GBel_J\, GBel_i\, \varphi$ is in general consistent in our logic. This means that where group beliefs are formed by principles other than unanimity—such as by majority voting in selection committees—then all group members are supposed to publicly adopt the group belief. They may however privately disagree with the outcome (which may also be due to the fact that they lied when they expressed their individual beliefs). For example, once the view of the set of ministers $J$ of some government has been decided then it becomes a group belief of $J$ that every member who had disagreed changes his mind and adopts the government view. Note also that in the special case where $J$ equals the singleton $\{i\}$, the conjunction $\neg GBel_i\, \varphi \wedge GBel_i\, GBel_i\, \varphi$ is inconsistent due to axiom ($\mathrm{D}_{GBel_i}$) for $GBel_i$ (as well as axiom (IN−)).

The axiom ($\mathrm{D}_{GBel_i}$) corresponds to the constraint 5, and is proper to individual belief. It expresses that individuals are always constituted groups.

Using (IN+), (IN−) and ($\mathrm{D}_{GBel_i}$) we can moreover show the following:

$$GBel_J\, \varphi \leftrightarrow GBel_i\, GBel_J\, \varphi \quad \text{if } i \in J \tag{4}$$

$$\neg GBel_J\, \varphi \leftrightarrow GBel_i\, \neg GBel_J\, \varphi \quad \text{if } i \in J \tag{5}$$

These theorems express that every agent is aware of what is believed (resp. not believed) by the set of agents he is member of.

It follows from axioms (IN+) and (IN−) that the modal axioms 4 and 5 (Chellas 1980) are provable in $\mathcal{GL}$ for every $GBel_J$ operator: these are therefore normal modal operators of type K45. Together with axiom ($D_{GBel_i}$) this means that the logic of individual belief is the standard doxastic logic $KD45_n$.

### Soundness and Completeness of $\mathcal{GL}$

A formula $\varphi$ is a theorem of logic $\mathcal{GL}$ if $\varphi$ is provable from the axioms of classical propositional logic together with ($K_{GBel_J}$), (IN+), (IN−), (UNA), (AGR) and ($D_{GBel_i}$), by means of the inference rules modus of ponens and ($RN_{GBel_J}$). Theoremhood of $\varphi$ in $\mathcal{GL}$ is noted $\vdash_{\mathcal{GL}} \varphi$.

The inference rule $RN_{GBel_J}$ and the axiom $K_{GBel_J}$ tell us that our logic is a normal modal logic. Each of the other axioms, i.e., IN+, IN−, UNA, AGR, $D_{GBel_i}$ have the syntactical form of so-called Sahlqvist axioms (Sahlqvist 1975). Therefore each of them has a corresponding semantical constraint on frames, viz. our constraints 1–5 making up the class of $\mathcal{GL}$ models. Then by Sahlqvist's general completeness result, our axiomatics constitutes a sound and complete axiomatization of the formulas that are valid in the class of $\mathcal{GL}$ models.

**Theorem 1.** *For every formula $\varphi$, $\models_{\mathcal{GL}} \varphi$ if and only if $\vdash_{\mathcal{GL}} \varphi$.*

We note that Sahlqvist's theorem also implies that the extension by any combination of IN+, IN−, UNA, AGR, $D_{GBel_i}$ of the basic normal modal logic (axiomatized by $RN_{GBel_J}$, $K_{GBel_J}$) is sound and complete w.r.t. the class of models obeying the corresponding constraints.

## 3.4 Some Invalid Formulas

Here are some properties that we have chosen to reject. In the sequel, $J'$ denotes a subset of a set of agents $J$.

**Not all sets of agents are groups.** Contrarily to individual belief, we do not consider that axiom D should be valid for group belief:

$$\nvDash \neg GBel_J \bot$$

In order for $\neg GBel_J \bot$ to hold $J$ should not simply be a set of agents but rather a constituted group. This is also related to our axiom (AGR): for example the formula $GBel_J\, \varphi \wedge GBel_J\, GBel_i \neg\varphi$ should be consistent, but should imply that $J$ is not a constituted group. The latter is due to the fact that the members of $J$ (publicly) disagree about what should be a common body of beliefs.

**Being a constituted group is not closed under subsets.** We have:

$$\nvDash \neg GBel_J \bot \to \neg GBel_{J'} \bot$$

in particular when $J' \subset J$. For example, consider the set $J = \{1, 2, \ldots, 11\}$ of 11 agents making up a football team: $J$ is a constituted group, but none of its subsets is so. More precisely, every agent in $\{1, 2, \ldots, 11\}$ identifies himself as a member of the group and recognizes $J$ as a group. This does not entail that $\{1, 2, \ldots, 10\}$ constitute a group. Indeed, it is not the case that every agent in $\{1, 2, \ldots, 10\}$ recognizes $\{1, 2, \ldots, 10\}$ as a group because we consider that ten players do not constitute a football team.

**Group belief does not imply subgroup belief.** We have:

$$\nvDash (\neg GBel_J \perp \wedge GBel_J\ \varphi) \rightarrow (\neg GBel_{J'} \perp \wedge GBel_{J'}\ \varphi)$$

in particular when $J' \subset J$. We reject this because group belief should not imply individual belief. More generally, a fraction of a big group might disagree with a group belief entertained by the whole group if the group belief was e.g. obtained by a majority vote. For example, in 2007 the US Democrats held the group belief that Obama was the best candidate for presidency while the subset of Clinton supporters disagreed. Non-validity of the implication already follows from non-validity of the preceding implication. We moreover have $\nvDash GBel_J\ \varphi \rightarrow GBel_{J'}\ \varphi$.

### *3.5 Example*

To illustrate our logic we model the example of Sect. 1.

1. Agent $i_1$ (privately) believes that $i_0$ is smart: $GBel_{i_1}\ smart_{i_0}$.
2. Agents $i_1$ and $i_2$ discuss and reach the consensus that $i_0$ is not smart: $GBel_{\{i_1,i_2\}} \neg smart_{i_0}$. (This follows from $GBel_{\{i_1,i_2\}} GBel_{i_1} \neg smart_{i_0} \wedge GBel_{\{i_1,i_2\}} GBel_{i_2} \neg smart_{i_0}$.)
3. Agent $i_3$ joins the conversation and they attain the consensus that $i_0$ is smart: $GBel_{\{i_1,i_2,i_3\}}\ smart_{i_0}$.

This illustrates that we might have consistent beliefs of nested constituted groups $\{i_1\} \subset \{i_1, i_2\} \subset \{i_1, i_2, i_3\}$ about propositions that change at each level of nesting:

$$\neg GBel_{i_1} \perp \wedge GBel_{i_1}\ smart_{i_0} \wedge \neg GBel_{\{i_1,i_2\}} \perp \wedge GBel_{\{i_1,i_2\}} \neg smart_{i_0} \wedge \\ \neg GBel_{\{i_1,i_2,i_3\}} \perp \wedge GBel_{\{i_1,i_2,i_3\}}\ smart_{i_0}$$

Fig. 2 contains a model of the situation after the interaction between $i_1$ and $i_2$, that is described by the formula $\varphi = GBel_{i_1}\ smart_{i_0} \wedge \neg GBel_{\{i_1,i_2\}} \perp \wedge GBel_{\{i_1,i_2\}} \neg smart_{i_0}$.

## 4 Discussion

We now discuss several properties of our logic of group belief $\mathcal{GL}$. We first compare $\mathcal{GL}$ with the logic of common belief. We then assess $\mathcal{GL}$ with respect to Gilbert's and Tuomela's non-reductionist theories of group belief as expounded in Sect. 2.2.

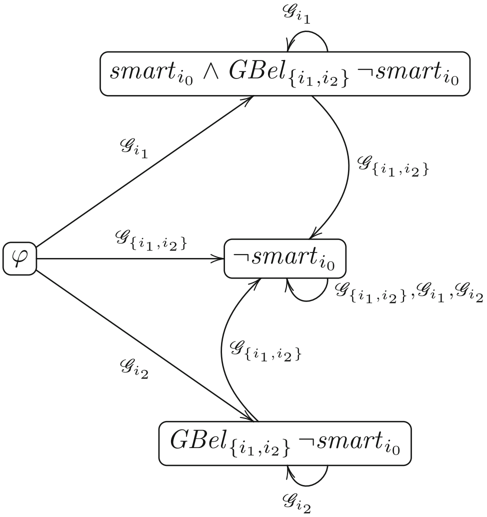

$\mathscr{G}_{i_1}$
$smart_{i_0} \wedge GBel_{\{i_1,i_2\}} \neg smart_{i_0}$
$\mathscr{G}_{i_1}$
$\mathscr{G}_{\{i_1,i_2\}}$
$\varphi$
$\mathscr{G}_{\{i_1,i_2\}}$
$\neg smart_{i_0}$
$\mathscr{G}_{\{i_1,i_2\}}, \mathscr{G}_{i_1}, \mathscr{G}_{i_2}$
$\mathscr{G}_{\{i_1,i_2\}}$
$\mathscr{G}_{i_2}$
$GBel_{\{i_1,i_2\}} \neg smart_{i_0}$
$\mathscr{G}_{i_2}$

**Fig. 2** Model after the interaction between agents $i_1$ and $i_2$ (Formula $\varphi$ holds in the left world)

In the third part we compare the logic of group belief and the logic of collective acceptance. We finally sketch a dynamic extension of $\mathcal{GL}$ where group beliefs can be updated, as in public announcement logic (PAL) (Baltag et al. 1998).

## 4.1 Discussion: The Relation Between Group Belief and Common Belief

Suppose we add a further modal operator $CBel_J$ to the language of $\mathcal{GL}$. The formula $CBel_J\ \varphi$ reads "the agents in $J$ commonly believe that $\varphi$ holds". Let us first recall semantics and axiomatics of common belief (Fagin et al. 1995).

### 4.1.1 Semantics of Common Belief

Common belief of a group of agents is semantically defined from individual belief: the mapping $\mathscr{C} : (2^{AGT} \setminus \{\emptyset\}) \longrightarrow (W \longrightarrow 2^W)$ associates an accessibility relation $\mathscr{C}_J$ to each group $J \subseteq AGT$ such that

6. $\mathscr{C}_J = (\bigcup_{i \in J} \mathscr{G}_i)^+$

For each group $J$, $\mathscr{C}_J$ is therefore the transitive closure of the union of the set of accessibility relations associated to $J$'s members. (Remember that we identify $GBel_i$ with the operator of individual belief.) $\mathscr{C}_J(w)$ is the set of possible worlds compatible with common beliefs of the group $J$.

So common belief is an aggregative kind of collective belief: it can be semantically reduced to individual beliefs (in terms of accessibility relations).[9]

### 4.1.2 Axiomatics of Common Belief

Axiomatically, common belief is defined by the Fixpoint Axiom ($\text{FP}_{CBel_J}$) and the Least Fixpoint Axiom ($\text{LFP}_{CBel_J}$):

$$CBel_J\,\varphi \leftrightarrow \bigwedge_{i\in J} GBel_i\,(\varphi \wedge CBel_J\,\varphi) \tag{$\text{FP}_{CBel_J}$}$$

$$\left(\bigwedge_{i\in J} GBel_i\,\varphi \wedge CBel_J\,(\varphi \rightarrow \bigwedge_{i\in J} GBel_i\,\varphi)\right) \rightarrow CBel_J\,\varphi \tag{$\text{LFP}_{CBel_J}$}$$

It follows from these axioms that the logic of $CBel_J$ contains KD4; in particular:

$$CBel_J\,\varphi \rightarrow \neg CBel_J\,\neg\varphi \tag{$\text{D}_{CBel_J}$}$$

$$CBel_J\,\varphi \rightarrow CBel_J\,CBel_J\,\varphi \tag{$4_{CBel_J}$}$$

Note that the negative introspection axiom 5 is not a theorem for $CBel$: the formula $\neg CBel_J\,\varphi \rightarrow CBel_J\,\neg CBel_J\,\varphi$ is not valid (Bonanno and Nehring 2000). In particular, from ($\text{D}_{CBel_J}$) and ($5_{CBel_J}$) we could deduce that $CBel_{J'}\,CBel_J\,\varphi \rightarrow CBel_J\,\varphi$ holds, which would mean that a group member cannot be wrong about a common belief of the group.

It follows from the Fixpoint Axiom that common belief implies individual belief:

$$CBel_J\,\varphi \rightarrow \bigwedge_{i\in J} GBel_i\,\varphi \tag{6}$$

### 4.1.3 Common Belief vs Group Belief

Now we establish a link between common belief and group belief.

**Proposition 1.** *The equivalence*

$$GBel_J\,\varphi \leftrightarrow CBel_J\,GBel_J\,\varphi$$

*is provable from the axioms of $\mathcal{GL}$ plus the axioms for common belief.*

[9] Note that this reduction has no syntactical counterpart: it would require an infinite conjunction. As both Gilbert and Tuomela use common knowledge and common belief operators, Gilbert's simple account is the only reductionist approach where collective beliefs can be syntactically reduced to individual beliefs.

*Proof.* The proof goes as follows:

1. $\vdash CBel_J\, GBel_J\, \varphi \rightarrow GBel_i\, GBel_J\, \varphi$, by ($\mathrm{FP}_{CBel_J}$)
2. $\vdash CBel_J\, GBel_J\, \varphi \rightarrow GBel_J\, \varphi$ from 1. by Theorem (4)
3. $\vdash GBel_J\, \varphi \rightarrow GBel_i\, GBel_J\, \varphi$ by Theorem (4), for every $i \in J$
4. $\vdash GBel_J\, \varphi \rightarrow \bigwedge_{i \in J} GBel_i\, GBel_J\, \varphi$ from 3.
5. $\vdash CBel_J\, (GBel_J\, \varphi \rightarrow \bigwedge_{i \in J} GBel_i\, GBel_J\, \varphi)$
   from 4. by the Rule of Necessitation for $CBel_J$
6. $\vdash CBel_J\, (GBel_J\, \varphi \rightarrow \bigwedge_{i \in J} GBel_i\, GBel_J\, \varphi) \rightarrow$
   $(\bigwedge_{i \in J} GBel_i\, GBel_J\, \varphi \rightarrow CBel_J\, GBel_J\, \varphi)$ from axiom $\mathrm{LFP}_{CBel_J}$
7. $\vdash \bigwedge_{i \in J} GBel_i\, GBel_J\, \varphi \rightarrow CBel_J\, GBel_J\, \varphi$ from 5. and 6. by Modus Ponens
8. $\vdash GBel_J\, \varphi \rightarrow CBel_J\, GBel_J\, \varphi$ from 4. and 7.
9. $\vdash GBel_J\, \varphi \leftrightarrow CBel_J\, GBel_J\, \varphi$ from 2. and 8.

■

Proposition 1 tells us that every group belief is commonly believed. Remember that this was one of the requirements for group belief of Sect. 2.3 (we will come back to that in Sect. 4.2). This property comes from the public nature of our operator$GBel_J$ , cf. the attention property mentioned in Sect. 3.2.

**Proposition 2.** *The equivalence*

$$(\neg GBel_J \perp \wedge GBel_J\, \varphi) \leftrightarrow CBel_J\, (\neg GBel_J \perp \wedge GBel_J\, \varphi)$$

*is provable from the axioms of $\mathcal{GL}$ plus the axioms for common belief.*

*Proof.* By Proposition 1, the schema $GBel_J\, \varphi \leftrightarrow CBel_J\, GBel_J\, \varphi$ is provable. Substituting $\varphi$ by $\neg GBel_J \perp \wedge \varphi$ we obtain the theorem

$$GBel_J\, (\neg GBel_J \perp \wedge \varphi) \leftrightarrow CBel_J\, GBel_J\, (\neg GBel_J \perp \wedge \varphi).$$

Then the proposition follows from the K45 theorem

$$GBel_J\, (\neg GBel_J \perp \wedge \varphi) \leftrightarrow (\neg GBel_J \perp \wedge GBel_J\, \varphi)$$

together with the rule of replacement of proved equivalents. ■

We highlight that contrarily to common belief, the negative introspection (axiom 5) holds for group belief. This comes from the fact that the public nature of group belief is stronger than that of common belief. Common belief is public in the sense that if a proposition is commonly believed then this common belief itself is commonly believed. More generally, we can say that a formula $\varphi$ is public for the group $J$ if and only if $\varphi \leftrightarrow CBel_J\, \varphi$. In this sense, any group belief is public (thanks to Proposition 1). But the publicness of the group belief is stronger because of axiom 5 (i.e., Axiom IN− with $J' = J$): if a group belief does not hold then the group believes (and thus is aware) that it does not hold. This makes that contrarily to common belief, individual belief cannot be wrong about group beliefs, as already observed in Sect. 4.1.2.

While it should be clear by now that contrarily to common belief, group belief does not logically imply individual belief, it may nevertheless be considered that it does so *by default*: when we learn that group $J$ believes that $\varphi$ then we often infer that $J$'s members individually believe that $\varphi$. Technically, this could be done by integrating nonmonotonic reasoning mechanisms into $\mathcal{GL}$.

In Sect. 4.4 we will present a dynamic variant of $\mathcal{GL}$. Some more differences between group belief and common belief will show up in that setting.

## 4.2 Back to the Philosophical Origins

We now revisit the criteria for group belief that we have put forward in Sect. 2.3 in the light of our logic.

### 4.2.1 Group Belief Features

**Group belief has a binding force.** As already said, this feature of group belief is made explicit in the reading of our belief operator. We have said that a situation where we have a genuine group belief by $J$ that $\varphi$ should be described by the formula $GBel_J\, \varphi \wedge \neg GBel_J\, \bot$: $J$ is a constituted group in Gilbert's sense, and the $J$ believes $\varphi$.

**No combination of individual beliefs implies group belief.** In our logic, there is no entailment link between individual beliefs and group beliefs: so in $\mathcal{GL}$, $\bigwedge_{i \in J} GBel_i\, \varphi$ neither implies $\neg GBel_J\, \bot$, nor does it imply $GBel_J\, \varphi$, so *a fortiori* $\bigwedge_{i \in J} GBel_i\, \varphi$ does not imply $\neg GBel_J\, \bot \wedge GBel_J\, \varphi$. We can generalize this proof to any combination of individual beliefs, and in particular to common belief.

**Group belief does not imply individual belief.** Group belief does not imply individual, private belief in our logic: for every agent $i$, be it a member of the group $J$ or not, $(\neg GBel_J\, \bot \wedge GBel_J\, \varphi) \rightarrow GBel_i\, \varphi$ is not valid in our logic.

**There is a commitment on group belief.** It is important to note that our logic does not have a separate operator of commitment from which group belief would be defined. Instead, we consider that our notion of group belief incorporates a notion of commitment. This property takes the form:

$$(\neg GBel_J\, \bot \wedge GBel_J\, \varphi) \rightarrow GBel_J \bigwedge_{i \in J} GBel_i\, \varphi$$

This formula is a theorem of our logic due to axiom (AGR) (actually even without the premiss $\neg GBel_J\, \bot$). A belief of group $J$ that $\varphi$ implies the commitment of each group member $i \in J$ to $\varphi$, in the sense that $i$ is declaring (implicitly and towards the group) that he believes $\varphi$. This is 'hard-wired' in the logic: a constituted group belief that $\varphi$ (i.e., $\neg GBel_J\, \bot \wedge GBel_J\, \varphi$) logically implies group belief that every member believes that $\varphi$ ($GBel_J\, GBel_i\, \varphi$, for $i \in J$).

Moreover, if an agent $i$ violates his commitment—e.g. by expressing a contrary point of view—then this destroys the group: the formula $GBel_J\, \varphi \wedge GBel_J\, GBel_i\, \neg\varphi \rightarrow GBel_J\, \bot$ is provable (thanks to AGR and $\mathrm{D}_{GBel_i}$). So the agent is committed to the group beliefs in the sense that, if he wants to stay member of a constituted group *Group* then he has to act according to beliefs of *Group*. (This will be made more explicit in our dynamic extension of $\mathcal{GL}$.)

**The group members share a common belief about group beliefs.** This is a theorem of our logic: as proved above, the formula $GBel_J\, \varphi \rightarrow CBel_J\, GBel_J\, \varphi$ is provable. Our logic is even stronger because we have an equivalence here.

To sum it up: our logic $\mathcal{GL}$ satisfies the list of the requirements for a group belief operator that we have postulated in Sect. 2.3. We now examine more deeply its link with Gilbert's and Tuomela's approaches.

### 4.2.2 Comparison with Gilbert's Plural Subject Account

In the sequel, we show that our group belief operator captures Gilbert's group belief definition of Gilbert (1989). This is mainly due to axioms (AGR) and (UNA). In particular, if we consider that $GBel_J\, GBel_i\, \varphi$ typically results from agent $i$ expressing in front of group $J$ that he believes $\varphi$ (and this fact being collectively accepted as a group belief) then Axiom (UNA): $GBel_J\, ((\bigwedge_{i\in J} GBel_i\, \varphi) \rightarrow \varphi)$ says that a group belief results from the expression of an individual belief by all members of the group. By making public their belief that $\varphi$ holds, the agents in $J$ publicly express their opinions that $\varphi$ should be accepted by the group $J$.

Moreover, from the theorem (3) of Sect. 3.3, Proposition 1 and the rule of substitution of proved equivalents we can deduce the equivalence:

$$GBel_J\, \varphi \leftrightarrow CBel_J\, (\bigwedge_{i\in J} GBel_J\, GBel_i\, \varphi) \tag{7}$$

According to the above remark this formula may be read: "$\varphi$ is a group belief of $J$ if and only if it is common belief in $J$ that every group member publicly expressed that he believes $\varphi$". This equivalence is thus very close to Gilbert's characterization of group belief.[10]

### 4.2.3 Comparison with Tuomela's Account

Our simple logical framework does not allow to capture the whole complexity of Tuomela's refinement of Gilbert definition of group belief that we have mentioned in the end of Sect. 2.2.1. In particular, in $\mathcal{GL}$ we do not have roles, institution or

[10] It may be argued that we do not entirely capture Gilbert's intended sense of commitment in our logic: a fully-fledged account should have a primitive modal operator of commitment. In any case, we believe that $\mathcal{GL}$ is the best that can be done with a logic that has only modal operators of belief.

norms, and we cannot distinguish between operative and non-operative agents. We however note that our logic can be easily extended by introducing the concept of '*leaders* of a group $J$ about a proposition $\varphi$', noted $leaders(J, \varphi)$. The leaders would be a subgroup of the group of agents $J$, verifying properties such as: $GBel_J\ (GBel_{leaders(J,\varphi)}\ \varphi \leftrightarrow \varphi)$. This means that it is grounded for the whole group $J$ that if it is grounded for its leaders that $\varphi$, then $\varphi$ true (i.e., if leaders have jointly accepted $\varphi$, then the other agents have to accept it tacitly and thus $\varphi$ becomes a proper group belief *à la* Tuomela). The group of leaders could be for example the government for every decision concerning the whole nation or a group of specialists of a domain for every fact concerning their domain of competence.

## 4.3 *The Relationships Between the Logic of Group Belief and the Logic of Acceptance*

The logic $\mathcal{AL}$ (*Acceptance Logic*) that was introduced in Gaudou et al. (2008), Lorini et al. (2009), and Herzig et al. (2009) has some similarities with our logic of group belief $\mathcal{GL}$. $\mathcal{AL}$ allows to express what agents accept while functioning as members of a certain institution $x$; in particular, $\mathcal{AL}$ allows to express that some agents identify themselves as members of $x$.

The logic $\mathcal{AL}$ has been exploited in order to model the relationship between acceptances and institutions and, in particular, in order to clarify how the existence and the dynamics of *norms* and *rules* of an institution might depend on their *acceptance* by the members of the institution. In the logic of acceptance, institutions are conceived as rule-governed social practices on the background of which the agents reason. For example, take the case of a game like Clue. The institutional context is the rule-governed social practice which the agents conform to in order to be competent players and on the background of which agents reason. In the context of Clue, an agent accepts that something has happened *qua* player of Clue (e.g., the agent accepts that Mrs. Red is the murderer *qua* player of Clue). The logic $\mathcal{AL}$ is aimed at capturing the state of acceptance *qua* member of an institution as the kind of acceptance one is committed to when one is "functioning as a member of the institution". It is proved in Lorini et al. (2009) that the logic of acceptance $\mathcal{AL}$ embeds the logic of normative systems of Grossi et al. (2006).

$\mathcal{AL}$ has operators for acceptance of the form $A_{J:x}$, which are interpreted by means of accessibility relations $\mathscr{A}_{J:x}$ between states in a model. The formula $A_{J:x}\varphi$ is read 'the agents in $J$ accept that $\varphi$ while functioning as members of institution $x$'. The formula $A_{J:x}\bot$ has therefore to be read 'agents in $J$ are not functioning as members of institution $x$'; conversely, $\neg A_{J:x}\bot$ has to be read 'agents in $J$ are functioning as members of institution $x$'. Thus, $\neg A_{J:x}\bot \wedge A_{J:x}\varphi$ means 'agents in $J$ accept that $\varphi$ *qua* members of institution $x$'. For singletons the formula $\neg A_{i:x}\bot \wedge A_{i:x}\varphi$ has to be read 'agent $i$ accepts that $\varphi$ *qua* member of institution $x$'.

The axiomatization of $\mathcal{AL}$ is as follows, where $x$ and $y$ denote institutional contexts.

$$\text{All K-principles for the operators } A_{J:x} \tag{K}$$

$$A_{J:x}\varphi \to A_{J':y}A_{J:x}\varphi \text{ for } J' \subseteq J \tag{IN$+_{A_{J:x}}$}$$

$$\neg A_{J:x}\varphi \to A_{J':y}\neg A_{J:x}\varphi \text{ for } J' \subseteq J \tag{IN$-_{A_{J:x}}$}$$

$$A_{J:x}\left(\left(\bigwedge_{i\in J} A_{i:x}\varphi\right) \to \varphi\right) \tag{UNA$_{A_{J:x}}$}$$

$$(\neg A_{J:x}\bot \wedge A_{J:x}\varphi) \to A_{J':x}\varphi \text{ for } J' \subseteq J \tag{INCL}$$

$$\neg A_{J:x}\bot \to \neg A_{J':x}\bot \text{ for } J' \subseteq J \tag{CS}$$

Axioms IN$+_{A_{J:x}}$ and IN$-_{A_{J:x}}$ are introspection axioms for acceptance which are similar to the Axioms IN+ and IN− for group belief. Axiom UNA$_{A_{J:x}}$ is a unanimity principle which describes the *bottom up* process leading from individual acceptances of the members of an institution to the collective acceptance of the group of members of the institution. This axiom is similar to Axiom UNA of the logic of group belief.

The Inclusion Axiom INCL says that, if the agents in $J$ accept that $\varphi$ *qua* members of $x$ then every subset $J'$ of $J$ accepts $\varphi$ while functioning as members of $x$. This means that things accepted by the agents in $J$ *qua* members of $x$ are necessarily accepted by the agents in all of $J$'s subsets with respect to the same institutional context $x$. The axiom describes the *top down* process leading from $J$'s collective acceptance to the individual acceptances of $J$'s members.

We observe that there is no such principle in $\mathcal{GL}$. In fact, an axiom such as $(\neg GBel_J \bot \wedge GBel_J \varphi) \to GBel_{J'} \varphi$ with $J' \subseteq J$ would be too strong: as we have argued, a group's beliefs may differ from the beliefs of its supergroups. On the contrary, Axiom INCL sounds reasonable for the logic of acceptance which mentions explicitly the (institutional, conversational or social) context in which the acceptance of a group of agents is taking place.

Another difference with $\mathcal{GL}$ is that in $\mathcal{AL}$, groups of an institution are supposed to be closed under subsets (Axiom CS). In fact, the formula $\neg A_{J:x}\bot$ is not aimed at capturing a strong notion of 'constituted group'. As said above, the $\mathcal{AL}$ formula $\neg A_{J:x}\bot$ has to be read 'agents in $J$ are functioning as members of $x$'. The latter expression just means that: every agent in $J$ identifies himself as a member of $x$ and recognizes every agent in $J$ as a member of $x$. Under this assumption, $\neg A_{J:x}\bot$ should imply $\neg A_{J':x}\bot$, for $J' \subseteq J$. As argued in Sect. 3.4, Axiom CS would be unreasonable for $\mathcal{GL}$, where the formula $\neg GBel_J \bot$ is aimed at capturing a notion of constituted group.

Although $\mathcal{GL}$ and $\mathcal{AL}$ have some different principles and properties, under certain conditions it is possible to find a translation from $\mathcal{GL}$ to $\mathcal{AL}$ such that all axioms and rules of inference of $\mathcal{GL}$ are theorems of $\mathcal{AL}$. To this end, a single institutional context $x_0$ is enough. Consider the following translation *tr* from $\mathcal{GL}$ to $\mathcal{AL}^+$:

- $tr(p) = p$,
- $tr(\neg\varphi) = \neg tr(\varphi)$,
- $tr(\varphi \vee \psi) = tr(\varphi) \vee tr(\psi)$,
- $tr(GBel_J\, \varphi) = A_{J:x_0} tr(\varphi)$.

**Proposition 3.** *For every $\mathcal{GL}$ formula $\varphi$, if $\varphi$ is a theorem of $\mathcal{GL}$ then*

$$\bigwedge\nolimits_{i \in AGT} \neg A_{i:x_0} \bot \models tr(\varphi)$$

*is a theorem of $\mathcal{AL}^+$ (where $\models$ is logical consequence with global axioms as defined in Sect. 3.2).*

Although Proposition 3 highlights some interesting relationships between $\mathcal{AL}$ and $\mathcal{GL}$, we cannot prove that the former embeds the latter, i.e. it seems difficult to find a straightforward translation *tr* from $\mathcal{GL}$ to $\mathcal{AL}$ such that $\varphi$ is $\mathcal{GL}$ satisfiable iff $tr(\varphi)$ is $\mathcal{AL}$ satisfiable. In fact, the two logics aim to capture different kinds of individual and collective attitudes. In Gaudou et al. (2008) and Lorini et al. (2009) the authors were interested in clarifying what 'accepting something *qua* member of an institution' means, and in studying the relationships between acceptances and institutions. In the present work, we have provided a logical formalization of group belief trying to account for the main properties of this concept which have been identified in the philosophical literature on collective Intentionality. Both works are however part of the same general program which consists in the logical analysis of different kinds of collective and group attitudes and of their relationships with individual attitudes (beliefs, preferences, etc.) and with social structures like institutions and organizations.

## 4.4 A Dynamic Variant of the Logic of Group Belief

In this section we discuss a dynamic variant of our logic where group beliefs are updated by *public announcements*. Our logic extends $\mathcal{GL}$ just as public announcement logic (PAL) extends epistemic logic (van Ditmarsch et al. 2007). We extend $\mathcal{GL}$ by modal operators of public announcement of the form $[\varphi!]$. $[\varphi!]\psi$ reads "if $\varphi$ is publicly announced then $\psi$ is true afterwards".

We adopt Kooi's semantics (which is a variant of the original PAL (Kooi 2007)) because it better suits belief (while the original PAL better suits knowledge). The truth condition for the operators of public announcement is:

$$M, w \models [\varphi!]\psi \;\text{ iff }\; M^{\varphi!}, w \models \psi$$

The components of the update $M^{\varphi!}$ of $M$ by $\varphi!$ are defined as follows:

- $W^{\varphi!} = W$;
- $u\mathcal{G}_J^{\varphi!}v$ iff $w\mathcal{G}_J v$ and $M, v \models \varphi$, for every $J \subseteq AGT$;
- $\mathcal{V}^{\varphi!}(p) = \mathcal{V}(p)$, for every $p \in ATM$.

The effect of a public announcement $\varphi!$ is to restrict the set of worlds that are compatible with what is believed by the group $J$ to the set of worlds in which $\varphi$ is true, for every group $J$. Note that it might be the case that before the announcement of $\varphi$, an agent $i$ believes that $\neg\varphi$: then the announcement empties $i$'s set of possible worlds. This is the reason why the semantic constraint 5 given in Sect. 3.2 must be abandoned in this dynamic extension of the logic of group belief. For the rest, it is straightforward to verify that if $M$ satisfies the semantic constraints 1–4 of in Sect. 3.2 then $M^{\varphi!}$ does so, too.

Call $\mathcal{GL}^-$ the variant of $\mathcal{GL}$ without constraint 5, and call PA-$\mathcal{GL}^-$ the extension of $\mathcal{GL}^-$ by public announcements. It is a routine task to prove that the following equivalences are valid in PA-$\mathcal{GL}^-$:

$$[\varphi!]p \leftrightarrow p \qquad (\text{Red}_p)$$

$$[\varphi!]\neg\varphi \leftrightarrow \neg[\varphi!]\varphi \qquad (\text{Red}_\neg)$$

$$[\varphi!](\psi \wedge \chi) \leftrightarrow ([\varphi!]\psi \wedge [\varphi!]\chi) \qquad (\text{Red}_\wedge)$$

$$[\varphi!]GBel_J\,\psi \leftrightarrow GBel_J\,(\varphi \rightarrow [\varphi!]\psi) \qquad (\text{Red}_{GBel_J})$$

They make up a complete set of reduction axioms: together with the rule of replacement of proved equivalences they allow to ‘push’ the dynamic operators $[\varphi!]$ through the logical operators of $\mathcal{GL}$, and in this way to reduce every formula containing dynamic operators to a provably equivalent formula without dynamic operators. So, completeness of PA-$\mathcal{GL}^-$ follows from the known completeness of the base logic without dynamic operators that we have established in Sect. 3.3 (more precisely, of the variant $\mathcal{GL}^-$ without the D-axiom, which can be proved straightforwardly).

The reduction axiom ($\text{Red}_{GBel_J}$) for group belief highlights an important difference between our logic and the logic of common belief: there is no such reduction axiom for common belief and common knowledge (Kooi and van Benthem 2004). Technically, this difference can be explained by the way group belief and common belief relate to individual beliefs. On the one hand, common belief is strongly linked to individual beliefs and can be semantically reduced to them: common belief of a set of agents $J$ is interpreted by means of the transitive closure of the union of the accessibility relations associated to the individuals in $J$. In contrast, the accessibility relation for group belief of $J$ cannot be defined from those for individual beliefs. In other words, a group belief of $J$ entertains a much weaker link with the individual beliefs of the agents in $J$ than common belief does. The difference is perhaps easier to understand with an example.

Let $M$ be a model with three worlds $W = \{w, v, u\}$ such that $\mathscr{V}(p) = \{w, v\}$ and $\mathscr{V}(q) = \{w, u\}$. Suppose also that $\mathscr{G}_i = \{(w, v), (v, v), (u, u)\}$, $\mathscr{G}_j = \{(w, w), (v, u), (u, u)\}$ and $\mathscr{G}_{\{i,j\}} = \{(w, u), (v, u), (u, u)\}$. By definition of $\mathscr{C}_{\{i,j\}}$, we have $\mathscr{C}_{\{i,j\}} = \{(w, w), (w, v), (w, u), (v, u), (v, v), (u, u)\}$. Note that even though $M, w \not\models CBel_{\{i,j\}}\,(q \rightarrow [q!]p)$, we still have $M, w \models [q!]CBel_{\{i,j\}}\,p$. Indeed, in the model $M^{q!}$ resulting from the announcement of $q$ we have $\mathscr{C}^{q!}_{\{i,j\}}(w) = \{w\}$. In words, it is not common belief that $q$ implies that $p$ is the case after the public announcement of $q$, but after the public announcement of $q$ it is common belief that $p$. That is, common belief may appear 'out of the blue': it was not foreseeable by the agents and just 'pops up'. Consider now group belief. We have $M, w \not\models GBel_{\{i,j\}}\,(q \rightarrow [q!]p)$, and also $M, w \not\models [q!]GBel_{\{i,j\}}\,p$. Indeed, in the model $M^{q!}$ resulting from the announcement of $q$ to $J$ we have $\mathscr{G}^{q!}_{\{i,j\}}(w) = \{w, u\}$. That is, contrary to common belief, group belief cannot just pop up if not previously foreseen by the agents.

Let us consider a further aspect of this dynamic extension of the logic of group belief. In Sect. 4.2.1 we have shown that if a (constituted) group $J$ believes that $\varphi$ (i.e., $\neg GBel_J \perp \wedge GBel_J\,\varphi$) then every member of the group is committed to the fact that $\varphi$ is true in front of the other members of the group (i.e., $GBel_J \bigwedge_{i\in J} GBel_i\,\varphi$): the group $J$ believes that each of its members might declare that he believes $\varphi$. Operations of public announcement can be used in order to model commitment dynamics and group belief dynamics by means of (very simple kind of) speech acts. The following valid formulas highlight this.

First,

$$[GBel_i\,p!]GBel_J\;GBel_i\,p \tag{8}$$

says that when $i$ asserts that $p$ is true publicly then $i$ becomes committed to the fact that $p$ is true towards $J$ (where $J$ might include $i$). We suppose indeed that the announcement $GBel_i\,p!$ captures a basic notion of assertion, that is, $GBel_i\,p!$ corresponds to the event "agent $i$ asserts that $p$ is true publicly".

Second,

$$[(\bigwedge_{i\in J} GBel_i\,\varphi)!]GBel_J\,p \tag{9}$$

says that after every agent in $J$ has asserted that $p$, $J$ starts to believe that $p$. Here is a sequential version, for $J = \{i_1, \ldots, i_n\}$:

$$[GBel_{i_1}\,\varphi!]\ldots[GBel_{i_n}\,\varphi!]GBel_J\,p \tag{10}$$

Let us prove theorem (9). First, observe that $[\bigwedge_{i\in J} GBel_i\,p!]GBel_J\,p$ is equivalent to

$$GBel_J\,((\bigwedge_{i\in J} GBel_i\,p) \rightarrow [\bigwedge_{i\in J} GBel_i\,p!]p)$$

(by the reduction axiom ($\mathrm{Red}_{GBel_J}$)). The latter is equivalent to $GBel_J\,((\bigwedge_{i\in J} GBel_i\,p) \rightarrow p)$ (by reduction axiom ($\mathrm{Red}_p$)), which is nothing but the Unanimity Axiom UNA. Therefore $[\bigwedge_{i\in J} GBel_i\,p!]GBel_J\,p$ must be a theorem of PA-$\mathcal{GL}^{-}$.

Before concluding, we note that our definition of assertion forces agents to trust the sincerity of other agents. Indeed, as a specific instance of formula (8) above, we have

$$[GBel_i\,p!]GBel_j\,GBel_i\,p \tag{11}$$

That is, when agent $i$ asserts $p$ then every agent $j$ believes that $i$ believes $p$. This is of course a limitation of this dynamic extension of $\mathcal{GL}$, one that that we intend to overcome in future works. To this aim, we plan to use an approach based on action/event models à la (Baltag et al. 1998). This amounts to defining an update operation which changes group beliefs without directly changing the agents' individual beliefs. Such an operation allows for situations where an agent $i$ asserts that $p$ towards a group $J$ and $J$ starts to believe that $i$ believes $p$ while no agent $j$ in $J$ individually believes that $i$ believes $p$ (as there is no agent $j$ trusting $i$). However, one has to be careful here: it is not straightforward to define an update operation in such a way that the updated model still satisfies the semantic constraints 1–5. For a solution in the case of acceptance see Herzig et al. (2009).

## 5 Conclusion

We have focussed on the characterization and the formalization of the notion of group belief, in the sense of a belief ascribed to a group as a whole. In the first part of the paper we have presented an overview of existing philosophical theories of collective belief (by presenting Gilbert's and Tuomela's accounts), and we have shown that a notion of group belief should not be confused with a reductionist view of collective belief as an aggregate of individual beliefs of some agents. Following this overview, we have highlighted the key features of group belief and have modeled them semantically and axiomatically in our logic $\mathcal{GL}$. We have then discussed the formal links between the notion of group belief (i.e., our group belief operator) and a reductionist notion of collective belief embodied by the common belief operator. In addition, we have presented a comparison between the logic of group belief and the logic of collective acceptance. Finally, we have discussed a dynamic variant of our logic, which is the first step to cover group belief formation.

The dynamic extension of Sect. 4.4 is only the first step to cover group belief formation as it occurs e.g. in Example 12. A more sophisticated account would require the integration of a theory of communication (typically, speech act theory (Searle 1969)) and of mechanisms of group belief formation beyond unanimity, as studied in social choice theory (Taylor 2005).

In the future we also plan to extend our formal analysis of collective attitudes to group intentions. To this end, we will have to supplement the logical framework presented in this paper with operators for individual goals (or individual preferences) of the form $Pref_i$ where $i$ refers to an individual agent and the formula $Pref_i\,\varphi$ means 'agent $i$ prefers $\varphi$' (or 'agent $i$ wants $\varphi$ to be true'). Again, following Gilbert, we can say that the agents in a group $J$ have the intention to do the joint action $\delta_J$ together (or the agents in $J$ are jointly ready to do the joint action $\delta_J$ together) if and only if the agents in $J$ have openly expressed their willingness to do $\delta_J$ together. We think that a preliminary formalization of this concept of group intention is expressed by the formula $GBel_J\,(\bigwedge_{i\in J} Pref_i\,\delta_J)$. We postpone to future work an in-depth analysis of this concept of group intention, of the formal relationships between the operator of group belief $GBel_J$, the operators of individual preference $Pref_i$, and the constructions for joint action of type $\delta_J$.

## References

Ågotnes, Thomas. 2012. What noone knows. In *10th conference on logic and the foundations of game and decision theory (LOFT 2012)*, Sevilla.

Aumann, Robert. 1976. Agreeing to disagree. *Annals of Statistics* 4: 1236–39.

Baltag, Alexandru, Larry Moss, and Slawomir Solecki. 1998. The logic of public announcements, common knowledge and private suspicions. In *Proceedings of the seventh conference on theoretical aspects of rationality and knowledge (TARK'98)*, ed. Itzhak Gilboa, 43–56. San Francisco: Morgan Kaufmann.

Bonanno, Giacomo, and Klaus Nehring. 2000. Common belief with the logic of individual belief. *Mathematical Logic Quarterly* 46(1): 49–52.

Bratman, Michael E. 1987. *Intention, plans, and practical reason*. Cambridge: Harvard University Press.

Brentano, Franz. 1995. *Psychology from an empirical standpoint*. London: Routledge.

Chellas, Brian F. 1980. *Modal logic: An introduction*. Cambridge/New York: Cambridge University Press.

Cohen, L. Jonathan. 1989. Belief and acceptance. *Mind* 391(XCVIII): 367–389.

Cohen, Philip R., and Hector J. Levesque. 1990. Intention is choice with commitment. *Artificial Intelligence Journal* 42(2–3): 213–261

Dennett, Daniel. 1987. *The intentional stance*. Cambridge: MIT.

Durkheim, Emile. 1982. *The rules of sociological method*. New York: Free.

Durkheim, Emile, and Marcel Mauss. 1963. *Primitive classification*. Chicago: University of Chicago Press.

Fagin, Ronald, Joseph Y. Halpern, Yoram Moses, and Moshe Y. Vardi. 1995. *Reasoning about knowledge*. Cambridge/London: MIT.

Gaudou, Benoit, Dominique Longin, Emiliano Lorini, and Luca Tummolini. 2008. Anchoring institutions in agents' attitudes: Towards a logical framework for autonomous multi-agent systems. In *International joint conference on autonomous agents and multiagent systems (AAMAS)*, Estoril, 728–735. ACM.

Gilbert, Margaret. 1987. Modelling collective belief. *Synthese* 73(1): 185–204.

Gilbert, Margaret. 1989. *On social facts*. London/New York: Routledge.

Gilbert, Margaret. 1996. *Living together: Rationality, sociality, and obligation*. Lanham: Rowman and Littlefield.

Gilbert, Margaret. 2002. Belief and acceptance as features of groups. *Protosociology* 16: 35–69.

Grossi, Davide, John-Jules Ch. Meyer, and Frank Dignum. 2006. Classificatory aspects of counts-as: An analysis in modal logic. *Journal of Logic and Computation* 16(5): 613–643.

Hakli, Raul. 2006. Group beliefs and the distinction between belief and acceptance. *Cognitive Systems Research* 7: 286–297.

Heal, Jane. 1978. Common knowledge. *Philosophical Quarterly* 28: 116–131.

Herzig, Andreas, Tiago de Lima, and Emiliano Lorini. 2009. On the dynamics of institutional agreements. *Synthese – Knowledge representation for agents and multi-agent systems* 171(2): 321–355.

Herzig, Andreas, Emiliano Lorini, Jomi F. Hübner, and Laurent Vercouter. 2010. A logic of trust and reputation. *Logic Journal of the IGPL* 18(1): 214–244. Special Issue "Normative Multiagent Systems".

Hintikka, Jaakko. 1962. *Knowledge and belief: An introduction to the logic of the two notions*. Ithaca: Cornell University Press.

Kooi, Barteld. 2007. Expressivity and completeness for public update logic via reduction axioms. *Journal of Applied Non-Classical Logics* 17(2): 231–253.

Kooi, B., and J. van Benthem. 2004. Reduction axioms for epistemic actions, In *AiML-2004: Advances in Modal Logic*, ed. R. Schmidt, I. Pratt-Hartmann, M. Reynolds, H. Wansing, number UMCS-04-9-1 in Technical Report Series, University of Manchester, pp. 197–211.

Lewis, David. 1969. *Convention*. Cambridge: Harvard University Press.

Lewis, David. 1972. Language and language. *Minnesota Studies for the Philosophy of Science* VII: 3–35.

Lorini, Emiliano, Dominique Longin, Benoit Gaudou, and Andreas Herzig. 2009. The logic of acceptance: Grounding institutions on agents' attitudes. *Journal of Logic and Computation* 19(6): 901–940.

Meijers, Anthonie. 1999. Believing and accepting as a group. In *Belief, cognition and the will*, 59–71. Tilburg: Tilburg University Press.

Meijers, Anthonie. 2002. Collective agents and cognitive attitudes. *Protosociology* 16: 20–85.

Meijers, Anthonie. 2003. Why accept collective beliefs? Reply to Gilbert. *Protosociology* 18–19: 377–388

Quinton, Anthony. 1976. Social objects. *Proceedings of the Aristotelian Society* LXXVI: 1–27.

Sahlqvist, Henrik. 1975. Completeness and correspondence in the first and second order semantics for modal logics. In *Proceedings of the 3rd Scandinavian Logic Symposium* (Univ. Uppsala, Uppsala, 1973), Studies in logic and the Foundations of Mathematics, vol. 82, ed. S. Kanger. Amsterdam: North-Holland

Schiffer, Stephen. 1972. *Meaning*. Oxford: Oxford University Press.

Searle, John R. 1969. *Speech acts: An essay in the philosophy of language*. New York: Cambridge University Press.

Searle, John R. 1983. *Intentionality: An essay in the philosophy of mind*. New York: Cambridge University Press.

Searle, John R. 1995. *The construction of social reality*. New York: Free Press.

Taylor, A.D. 2005. *Social choice and the mathematic of manipulation*. Cambridge/New York: Cambridge University Press.

Tollefsen, Deborah Perron. 2002. Challenging epistemic individualism. *Protosociology* 16: 86–117.

Tuomela, Raimo. 1992. Group beliefs. *Synthese* 91(3): 285–318.

Tuomela, Raimo. 2000. Belief versus acceptance. *Philosophical Explorations* 2: 122–137

van Ditmarsch, Hans, Wiebe van der Hoek, and Barteld Kooi. 2007. *Dynamic epistemic logic. Synthese Library*, vol. 337. Dordrech: Springer.

Wray, K. Brad. 2001. Collective belief and acceptance. *Synthese* 3(129): 319–333.

Wray, K. Brad. 2003. What really divides Gilbert and the rejectionnists. *Protosociology* 18–19: 363–376

# Logic of Promotion and Demotion

**Patrick Girard**

**Abstract** In a logic with a dimension that represents social networks, for example friendship, it is natural to add hierarchies. We can then talk about friends being better than others, and isolate best friends. However, hierarchies are not rigid: majors can become lieutenant, friendship may be strengthened or compromised, and experts can loose or gain credibility. A proper analysis of the dynamics of hierarchies is thus essential to the logic of social networks. Hierarchies of agents are structurally very similar to plausibility orders of possible worlds central to logics for belief dynamics. I use this formal analogy to show how standard policies of belief revision can be applied in social networks, thus providing systematic mechanisms of promotion and demotion in social networks.

**Keywords** Social networks • Dynamic logic • Belief revision • Logic in the community • Two-dimensional logic

What does promotion have to do with belief revision? Think of belief revision as dynamics over hierarchies of possible worlds. To revise with information $\varphi$ is to systematically promote worlds described by $\varphi$. If you now think of $\varphi$ as describing a group of agents, the $\varphi$-agents, then belief revision provides policies to systematically promote the $\varphi$-agents. Johan van Benthem (2007) describes the belief revision operations of *lexicographic upgrade* and *elite change*. About lexicographic upgrade, van Benthem says: "This move is like a social revolution where some underclass *P* now becomes the upper class." About elite change, he says: "Macchiavellistically,

I would like to thank Shaun White, Marcus Triplett and the anonymous referee for comments that improved the paper greatly.

P. Girard (✉)
University of Auckland, New Zealand
e-mail: p.girard@auckland.ac.nz

A. Herzig, E. Lorini (eds.), *The Cognitive Foundations of Group Attitudes and Social Interaction*, Studies in the Philosophy of Sociality 5,
DOI 10.1007/978-3-319-21732-1_5

one just co-opts the leaders of the underclass, leaving the further social order unchanged." Transferred to a social setting, elite change and lexicographic upgrade have a literal reading instead of an analogical interpretation. This idea is at the core of the logic of promotion and demotion.

I addressed the problem of promotion and demotion in Girard (2011) and Girard and Seligman (2009) with a logic for aggregation of prioritised preference orders (cf., Andréka et al. 2002). I used a logical language with modalities $[G]\varphi$ defined over the aggregated preferences of groups of agents $G$. For instance, I defined the modality $[i/j]\varphi$ over the aggregation of the preferences of agents $i$ and $j$ by giving priority to the preferences of agent $i$. I then analysed promotion as a shift from a group $G$ to a new group $i/G$ in which agent $i$ is given priority over other agents in $G$. Using this logical language, I could formalise the aggregated preferences over groups but I couldn't reason directly about the structure of the groups.

In this paper, I will propose a logic of promotion and demotion (LPD henceforth) building on the framework of Logic in the Community (cf., Seligman et al. 2011, 2013). Logic in the community is a two-dimensional logic with epistemic and social dimensions. The social dimension contains social networks: groups of agents socially related, for example by a relation of friendship $F$. The modal language for this logic has a corresponding *friendship* modality $\langle F\rangle\varphi$, allowing to express social statements like "Carol is my friend" by $\langle F\rangle$**Carol**. LPD adds to this framework hierarchy relations $H_a$ for each agent $a$. Hierarchies are simply total preorders over sets of friends. The language of LPD contains two modalities $\langle H_a\rangle\varphi$ and $\langle H_a^<\rangle\varphi$ defined over the hierarchy of $a$'s friends. You can read $\langle H_a\rangle\varphi$ as "$\varphi$ holds for some friend that is at least as good as", and $\langle H_a^<\rangle\varphi$ as "$\varphi$ holds for some better friend." For the dynamics of promotion and demotion, I use propositional dynamic logic (PDL, cf., Harel et al. 2000). As shown in Girard and Rott (2014), several belief revision policies are definable in PDL. In LPD, these are adapted to the social dimension, yielding various policies of promotion and demotion.

## 1 Hierarchical Models

Hierarchical models combine epistemic and social components in a two-dimensional framework. In the first dimension, possible worlds are ordered by agents according to *indistinguishability*. In the second dimension, there are two components: (1) a social network for each possible world, and (2) a hierarchy over each agent's friends. Propositions are evaluated at world-agent pairs. So you may think of propositions as being doubly indexical: *$p$ is true at world $w$ for agent $a$*.

Given a set of propositional variables PROP and agent names AGENT, *hierarchical models* are tuples $M = \langle W, A, K, F, H, H^<, V\rangle$, in which:

- $W$ is a non-empty set of possible worlds,
- $A = \{\underline{a}, \underline{b} \ldots\}$ is a finite set of agents,
- $K$ is an epistemic (equivalence) relation over $W \times A$ such that $\langle(w, \underline{a}), (v, \underline{b})\rangle \in K$ implies that $\underline{a} = \underline{b}$,
- $F$ is a *friendship*[1] relation over $W \times A$ such that $\langle(w, \underline{o}), (v, \underline{b})\rangle \in F$, implies that $w = v$,
- $H$ is a collection of total preorders[2] $H_{\underline{a}}$ on the set $\{(w, \underline{b}) \in W \times A \mid \underline{a} = \underline{b}$ or $\langle(w, \underline{a}), (w, \underline{b})\rangle \in F\}$ for every $\underline{a} \in \mathsf{AGENT}$ such that: $\langle(u, \underline{b}), (v, \underline{c})\rangle \in H_{\underline{a}} \Rightarrow u = v$,
- $H^<$ is a collection of strict orders $H^<_a$ defined as sub-relations of $H_a$ in the usual way[3]: $\langle(w, a), (v, b)\rangle \in H^<_a$ iff $\langle(w, a), (v, b)\rangle \in H_a$ and $\langle(v, b), (w, a)\rangle \notin H_a$, and
- $V$ is a propositional valuation which assigns subsets of $W \times A$ to propositional variables. To each agent name $a \in \mathsf{AGENT}$, $V$ assigns a unique agent $\underline{a} \in A$. So for $a \in \mathsf{AGENT}$, $V(a) = W \times \{\underline{a}\}$.

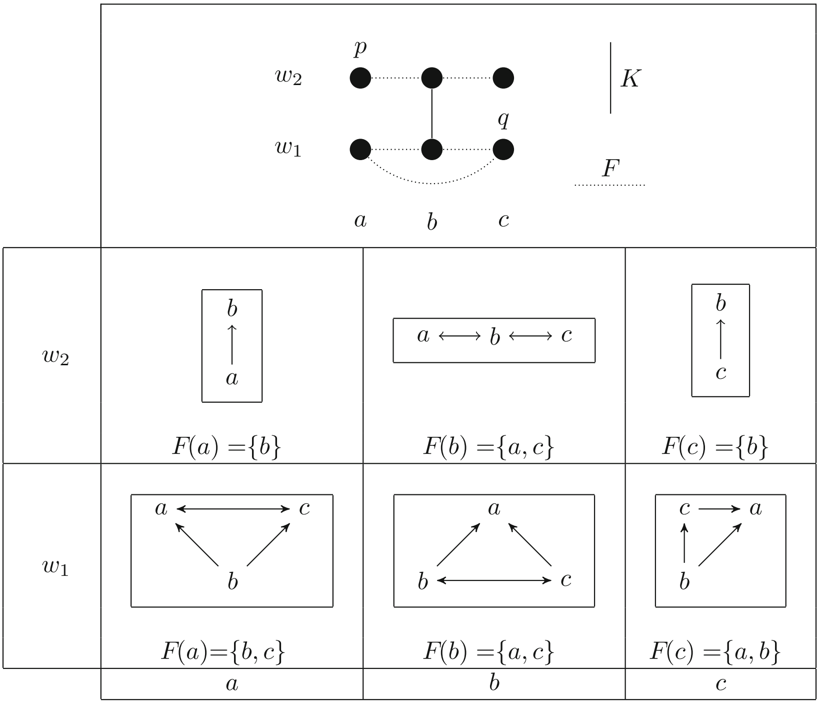
I will abuse notation and write $a$ indiscriminately to refer to agent names $a \in \mathsf{AGENT}$ or proper agents $\underline{a} \in A$. In hierarchical models, each agent has an epistemic relation over the set of possible worlds, and each world has a friendship relation over the set of agents. The domain of a hierarchical relation $H_a$ is the set of world-agent pairs $(w, b)$ such that $a$ and $b$ are friends in world $w$, and hierarchies are kept world-bound. So in each world and for any two friends, agents can tell whether they are equal friends, or if one is better than the other. If $\langle(w, b), (w, c)\rangle \in H_a$, say that "$c$ is at least as good a friend to $a$ as $b$". If $\langle(w, b), (w, c)\rangle \in H^<_a$, say that "$c$ is a better friend to $a$ than $b$".

*Example.* The following represents a hierarchical model, call it $M$. I will refer back to $M$ several times in the paper.

[1] I use friendship as a basic social relation for simplicity. I thus only assume $F$ to be symmetric. Other social relations could be used, but friendship is all I need for the interpretations of promotion and demotion I have in mind.

[2] Preorders are reflexive and transitive relations. Total preorders make any two friends comparable. Friends may be equally ranked, as you should expect.

[3] Because it is defined in terms of $H$, $H^<$ is redundant in models. But it is not redundant in the logic, as it is well-known that strict subrelations are not modally definable. For uniformity, I thus keep $H^<$ in models.



*M* is a two-dimensional model with two worlds, $w_1$ and $w_2$, and three agents, $a$, $b$ and $c$.[4] The top part represents the epistemic and friendship relations. For each world, there is a friendship relation represented with dotted horizontal lines. Hence, in $w_1$, all agents are friends together. In $w_2$, $b$ is friends with $a$ and $c$, but $a$ and $c$ are not friends. The vertical lines represent epistemic relations, and only agent $b$ finds worlds $w_1$ and $w_2$ indistinguishable. Since $a$ and $c$ are friends in $w_1$, but not in $w_2$, the model depicts a situation in which agent $b$ doesn't know whether $a$ and $c$ are friends. Finally, the proposition $p$ is true at $w_2$ for agent $a$ and $q$ is true at $w_1$ for agent $c$. The bottom part represents every agent's hierarchy over their friends. In $w_2$, $b$ ranks no one above others, but $a$ and $c$ rank $b$ above themselves. In $w_1$, $a$ ranks herself and $c$ equally above $b$, $b$ ranks $a$ above both $b$ and $c$, and finally $c$ puts $a$ on top of herself, with $b$ at the bottom.

## 2 Basic Language and Semantics

Let $\rho \in \mathsf{PROP} \cup \mathsf{AGENT}$. The basis of the $\mathsf{LPD}$-language for the logic of promotion and demotion is constructed from the following syntactic rules:

[4] Here and throughout the paper, I omit transitive and reflexive links whenever it improves readability in pictures.

$$\begin{aligned}\pi &::= K \mid F \mid H_a \mid H_a^< \\ \varphi &::= \rho \mid \neg\varphi \mid \varphi \wedge \varphi \mid \langle\pi\rangle\varphi\end{aligned}$$

The interpretation of the languages is an extension of the valuation function to a valuation $[\![\cdot]\!]^M$ assigning semantic values, or subsets of $W \times A$, to the sentences of the language.[5] In each hierarchical model $M = \langle W, A, K, F, H, H^<, V\rangle$, semantic values $[\![\varphi]\!]^M \subseteq W \times A$ and $[\![\pi]\!]^M \subseteq (W \times A)^2$ are computed in the following way:

$$\begin{aligned}
[\![\rho]\!]^M &= V(\rho), \text{ for } \rho \in \mathsf{PROP} \cup \mathsf{AGENT}. \\
[\![\neg\varphi]\!]^M &= W \setminus [\![\varphi]\!]^M \\
[\![\varphi \wedge \psi]\!]^M &= [\![\varphi]\!]^M \cap [\![\psi]\!]^M \\
[\![\langle\pi\rangle\varphi]\!]^M &= \{(w,a) \in W \times A \mid \langle(w,a),(v,b)\rangle \in [\![\pi]\!]^M \;\&\; (v,b) \in [\![\varphi]\!]^M, \\
&\quad \text{for some } (v,b) \in W \times A\} \\
[\![K]\!]^M &= K \\
[\![F]\!]^M &= F \\
[\![H_a]\!]^M &= H_a \\
[\![H_a^<]\!]^M &= H_a^<
\end{aligned}$$

*Example (continuing from p. 109).* Here are some formulas that are true in $M$.

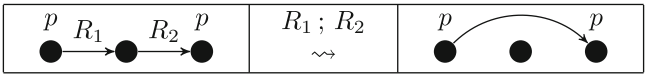
| | |
|---|---|
| $(w_1, b) \in [\![\neg[K]\langle H_b\rangle Kq]\!]^M$ | In world $w_1$, $b$ doesn't know that she has a friend at least as good as herself who knows $q$. |
| $(w_2, c) \in [\![[H_c^<]\neg[K]\langle F\rangle Kp]\!]^M$ | None of $c$'s better friends know that they have a friend who knows $p$. |
| $(w_2, b) \in [\![\langle K\rangle\langle H_c^<\rangle c]\!]^M$ | It is consistent with what $b$ knows that $c$ may rank herself above $b$. |

## 3 PDL Programs

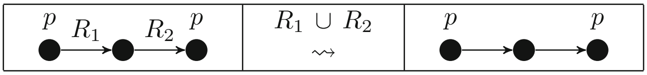
PDL-programs are tools for transforming models by redefining the relations between worlds using propositional dynamic logic (PDL). The new relations are constructed out of the old ones using PDL-*programs*. PDL-programs are built using four basic operations: *composition*, *choice*, *iteration* and *test*. From now on, I will only write 'program' instead of 'PDL-program'.

[5] I choose this notation for the definition of the semantics over the more common $M, w, a \models \varphi$ for uniformity and easier integration of PDL in the next sections. In the more common notation, instead of writing $(w, a) \in [\![\langle\pi\rangle\varphi]\!]^M$, we would write $M, w, a \models \langle\pi\rangle\varphi$.

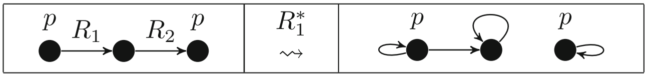
The *composition* program ';' takes two relations $R_1$ and $R_2$ and combines them so that $\langle x, y\rangle \in (R_1 \,;\, R_2)$ whenever there is a $z$ such that $R_1xz$ and $R_2zy$:



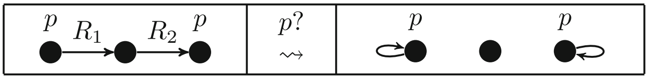
The *choice* program '∪' chooses between two relations $R_1$ and $R_2$ so that $\langle x, y\rangle \in (R_1 \cup R_2)$ if either $R_1xy$ or $R_2xy$:



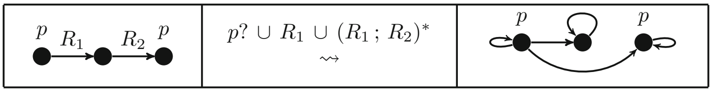
The *iteration* program '∗' repeats a basic program an arbitrary finite number of times. Formally, it corresponds to taking the reflexive transitive closure of a relation, as in:



Finally, the *test* program '?' tests if a formula is true at a state. As composition and choice, the test program defines a relation on models. It returns a reflexive link for worlds in which the tested formula is true[6]:



PDL can be used to define complex PDL programs. For example:



To describe programs in the language, we simply add the PDL-operators:

$$\pi ::= K \mid F \mid H_a \mid H_a^< \mid \pi \cup \pi \mid \pi \,;\, \pi \mid \pi^* \mid \varphi?$$
$$\varphi ::= \rho \mid \neg\varphi \mid \varphi \wedge \varphi \mid \langle\pi\rangle\varphi$$

And we expand the semantic definition accordingly:

[6] Maybe not very intuitive, but that's how it works.

$$\begin{array}{rcl}
[\![\pi_1 \cup \pi_2]\!]^M & = & [\![\pi_1]\!]^M \cup [\![\pi_2]\!]^M \\
[\![\pi_1 ; \pi_2]\!]^M & = & \{\langle (w,a),(v,b)\rangle \ \mid \ \langle (w,a),(u,c)\rangle \in [\![\pi_1]\!]^M \\
 & & \ \&\ \langle (u,c),(v,b)\rangle \in [\![\pi_2]\!]^M, \text{for some } (u,c) \in W \times A\} \\
[\![\pi^*]\!]^M & = & \{\langle (w,a),(v,b)\rangle \mid (w,a) = (v,b) \text{ or } \langle (u_i,a_i),(u_{i+1},a_{i+1})\rangle \\
 & & \in [\![\pi]\!]^M \text{ for some } n \geq 0, (u_0,a_0), \ldots (u_n,a_n) \in W \times A, \\
 & & (u_0,a_0) = (w,a) \text{ and } (u_n,a_n) = (v,b)\} \\
[\![\varphi ?]\!]^M & = & \{\langle (w,a),(w,a)\rangle \ \mid \ (w,a) \in [\![\varphi]\!]^M\}
\end{array}$$

## 4 Promotion and Demotion

Having a social language, we can describe groups of agents with formulas. For instance, we can isolate the friends of **Barry** and **Carol** with the formula $\langle F\rangle\mathbf{Barry} \vee \langle F\rangle\mathbf{Carol}$. For any world $w$, any formula $\varphi$ describes a group of agents, viz., the agents $a$ such that $M, w, a \models \varphi$. Hence, we can use belief revision operations on $\varphi$ to promote or demote groups of agents. I will use the following abbreviations:

$$\begin{array}{ll}
H_a^{\varphi} & ::= (\varphi ? \,;\, H_a \,;\, \varphi ?) \\
\mathbf{best_a}(\boldsymbol{\varphi}) & ::= \langle F\rangle a \wedge \varphi \wedge \neg\langle H_a^{<}\rangle\varphi
\end{array}$$

For any formula $\varphi$, $H_a^{\varphi}$ restricts $a$'s hierarchy to agents described by $\varphi$ and $\mathbf{best_a}(\boldsymbol{\varphi})$ isolates the best $\varphi$-agents in $a$'s hierarchy. For example, take $\varphi = \langle F\rangle a$, i.e., agents satisfying the formula which says that $a$ is amongst their friends, then $\mathbf{best_a}(\boldsymbol{\varphi})$ returns $a$'s best friends. Or one can think of $\varphi$ as ascribing expertise to agents, so that promoting $\varphi$-agents is giving priority to $\varphi$-experts.

I first consider two operations of promotion which I call, following the terminology of Girard and Rott (2014) and Rott (2009), *conservative* and *moderate*. Conservative promotion promotes the best $\varphi$-agents on top of the hierarchy and preserves the ranking otherwise:

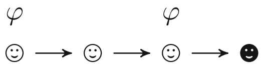
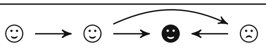
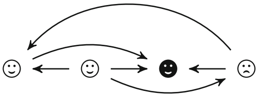
| Conservative Promotion |
|---|
| $\mathsf{CProm}_{\mathsf{a}}(\varphi) \quad = \quad H_a^{\neg\mathbf{best_a}(\boldsymbol{\varphi})} \cup ((\langle F\rangle a \wedge \neg\mathbf{best_a}(\varphi))? \,;\, F^* ; \, \mathbf{best_a}(\varphi)?)$ |

Notice the role of $F^*$ to ensure that all of $a$'s friends can be accessed, creating (possibly) new links ranking $a$'s best $\varphi$-friends over the others. Since $F$ is a symmetric relation, $F^*$ is an equivalence relation (it takes the reflexive transitive closer of a symmetric relation). Whenever I need to access all of $a$'s friends in programs, I use $F^*$ in a similar fashion.

Moderate promotion acts like conservative promotion, but promotes *all* $\varphi$-agents instead of only the best ones:

| Moderate promotion |
|---|
| $\mathsf{MProm_a}(\varphi) \quad = \quad H_a^{\varphi} \cup H_a^{\neg\varphi} \cup (((\langle F\rangle a \wedge \neg\varphi)? \,;\, F^* \,;\, (\langle F\rangle a \wedge \varphi)?)$ |

As a simple representation, here's the result of applying conservative and moderate promotion to the same initial model:

| Conservative and Moderate Promotion | | |
|---|---|---|
|  | $\mathsf{CProm_a}(\varphi)$ |  |
| | $\mathsf{MProm_a}(\varphi)$ |  |

The black figures represent best friends. The three operations of promotion agree on who should be the best friends after promoting $\varphi$ agents, but they disagree on how to order the remaining friends. Conservative promotion preserves most of the initial hierarchy, only taking the best $\varphi$-agents and putting them on top. Moderate promotion reorders every agent, by putting all $\varphi$-agents over all $\neg\varphi$-agents.

For demotion, I also define a conservative and a moderate version. As these operations are based on doxastic operations with a minimalist attitude, the result of demoting $\varphi$-agents doesn't entail that $\varphi$-agents are no longer best friends. What demotion does to a group is to make sure that the set of best friends is no longer only constituted by $\varphi$-agents.

Conservative demotion takes the best $\neg\varphi$-agents and puts them on a par with other best friends, but preserves the hierarchy otherwise.

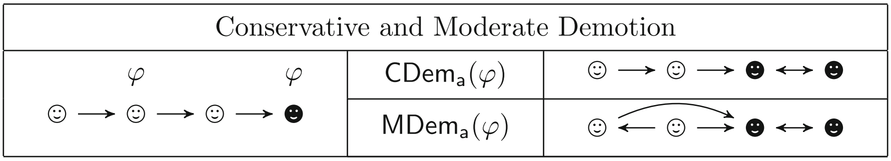
| Conservative Demotion |
|---|
| $\mathsf{CDem_a}(\varphi) \quad = \quad H_a^{\neg\mathbf{best_a}(\neg\boldsymbol{\varphi})} \cup (F^* \,;\, \mathbf{best_a}(\neg\boldsymbol{\varphi})?) \cup (F^* \,;\, \mathbf{best_a}(\top)?)$ |

Conservative demotion guarantees that the ruling class no longer consists only of $\varphi$-agents.

Moderate demotion again preserves best $\varphi$-friends, but it puts all other $\varphi$-agents under $\neg\varphi$-agents:

| Moderate Demotion |
|---|
| $\mathsf{MDem_a}(\varphi) \quad = \quad H_a^{\varphi} \cup H_a^{\neg\varphi}$<br>$\cup\, (((\langle F\rangle a \wedge \varphi \wedge \neg\mathbf{best_a}(\top))? \,;\, F^* \,;\, (\langle F\rangle a \wedge \neg\varphi)?)$<br>$\cup\, (F \,;\, \mathbf{best_a}(\neg\boldsymbol{\varphi})?) \cup (F^* \,;\, \mathbf{best_a}(\top)?)$ |

The following diagram illustrates the difference between conservative and moderate demotion. As was the case with promotion, the two operations agree on who become the best friends after demotion, but diverge in how they treat other agents.



## 5 PDL-Transformations

The final installments in the logic of promotion and demotion are PDL transformations, taken from Girard et al. (2012).[7] PDL-transformations are collections of PDL programs that operate *in parallel*. A PDL-transformation $\Lambda$ is a collection of programs $\Lambda(K)$, $\Lambda(F)$ and $\Lambda(H_a)$ that redefines each of the relations. I represent PDL-transformations in the following way:

| $\Lambda$ | | |
|---|---|---|
| $K$ | $:=$ | $\Lambda(K)$ |
| $F$ | $:=$ | $\Lambda(F)$ |
| $H_a$ | $:=$ | $\Lambda(H_a)$, for every $a \in$ AGENT |

A PDL-transformation is thus a way of combining several programs to redefine the relations of a model. From now on, I will just write 'transformation' instead of 'PDL-transformation'.

Let $\Lambda$ be a transformation and let $M = \langle W, A, K, F, H, H^{<}, V\rangle$ be a hierarchical model. $\Lambda(M) = \langle W, A, \Lambda(K), \Lambda(F), \Lambda(H_a), \Lambda(H_a^{<}), V\rangle$ is a new hierarchical model resulting from applying $\Lambda$ to $M$, in which:

$$\begin{array}{lcl} \Lambda(K) & = & [\![\Lambda(K)]\!]^M \\ \Lambda(F) & = & [\![\Lambda(F)]\!]^M \\ \Lambda(H_a) & = & [\![\Lambda(H_a)]\!]^M \\ \Lambda(H_a^{<}) & = & [\![(\Lambda(H_a))^{<}]\!]^M \end{array}$$

In some cases, transformations preserve some relations in the model exactly as they were. For instance, in the logic of promotion and demotion, they never affect

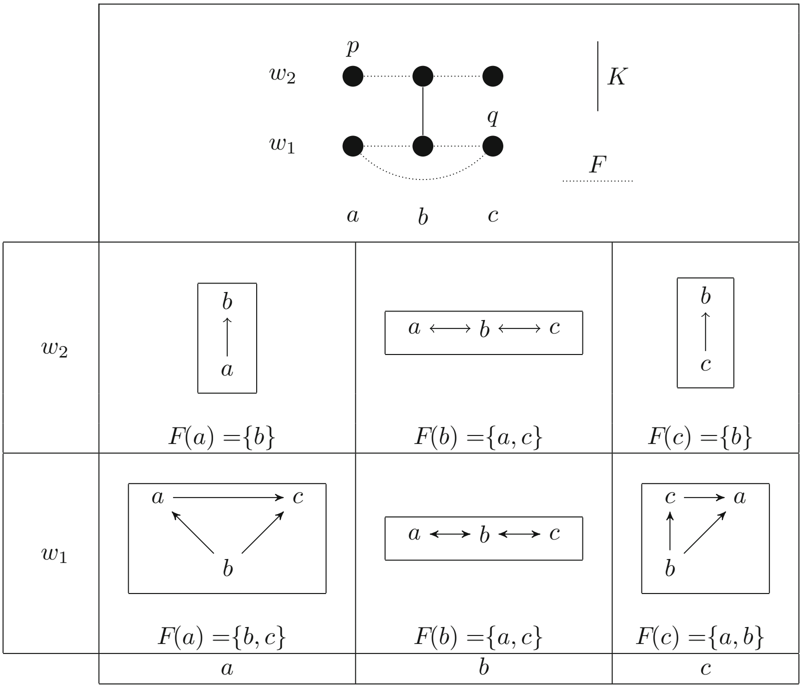
[7] For the details of the general case of PDL-transformations, the reader should consult section 1 of Girard et al. (2012). I give here a self-contained special case of PDL-transformations required for my purposes.

the epistemic relation. I will thus shorten the representation of transformations by omitting relations that are preserved, as in:

| $\beta$ | | |
|---|---|---|
| $H_a$ | $:=$ | $\mathsf{MProm_a}(c)$ |
| $H_b$ | $:=$ | $\mathsf{CDem_b}(a)$ |

The transformation $\beta$ contains two programs, one for $a$ and one for $b$. With $\beta$, $a$ moderately promotes $c$ and his friends, and $b$ conservatively demotes $a$. All other relations are not affected by $\beta$, and so are omitted from the representation.

*Example (continuing from p. 111).* Let's see how $\beta$ operates on $M$ by computing $\beta(M)$:

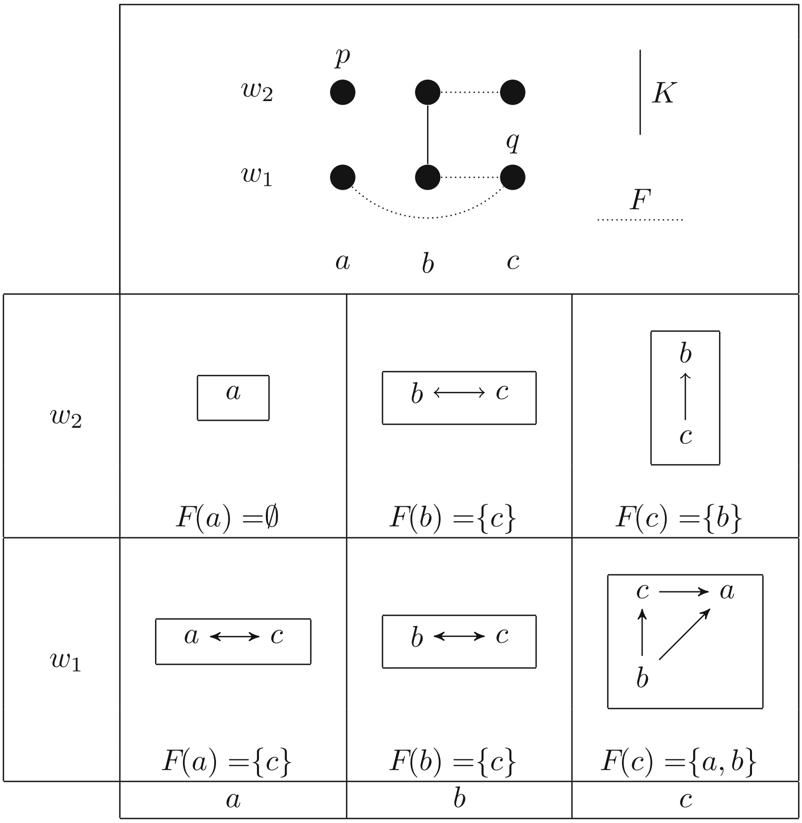


Transformations can do more than simply combining social action for all agents, as in the simple example above. They can also define actions of promotion and demotion that are not reducible to simple programs. As an example, here's a transformation for *radical* promotion that operates on both the friendship and the hierarchy relation to define a new hierarchy. Radical promotion is an operation by which an agent breaks the relationship with some of her friends, but keeps the hierarchy amongst the remaining friends as it was before. Since friendship is a symmetric relation, $a$ breaking the relationship with $b$ forces $b$ to also reconfigure her hierarchy:

| Radical Promotion – $\mathsf{RProm_a}(\varphi)$ | | |
|---|---|---|
| $F$ | $:=$ | $F^{\neg a} \cup (a?\,;F\,;\varphi?) \cup ((\neg a \wedge \varphi)?\,;F\,;a?)$ |
| $H_a$ | $:=$ | $H_a^{\varphi} \cup (a?\,;H_a\,;\varphi?) \cup (\varphi?\,;H_a\,;a?)$ |
| $H_i$ | $:=$ | $(\varphi?\,;H_i^{\neg a}) \cup (\neg\varphi?\,;H_i)$ for $i \neq a, i \in \mathsf{AGENT}$ |

This definition is tailored to the friendship relation $F$ being a symmetric relation. When $a$ drops $\varphi$-agents amongst her friends, those agents are no longer friends with $a$ and must adapt their hierarchy relations accordingly. $H_a$ is transformed so that $a$ only ranks herself and $\varphi$-agents just as she used to rank them. Finally, all other agents, if they are $\varphi$-agents, have to exclude $a$ from their hierarchies, as they are no longer friends.

*Example (continuing from p. 111).* As an example of a transformation acting on a model, let's compute the result $\mathsf{RProm_b}(c)(M)$ of $b$ radically promoting $c$ in model $M$:



# 6 Logic of Promotion and Demotion: LPD

To describe transformations in the language of LPD, we add modalities $\langle\Lambda\rangle$ for each acceptable transformation $\Lambda$[8]:

$$\begin{aligned} \pi &::= K \mid F \mid H_a \mid H_a^< \mid \pi \cup \pi \mid \pi\, ;\, \pi \mid \pi^* \mid \varphi? \\ \varphi &::= \rho \mid \neg\varphi \mid \varphi \wedge \varphi \mid \langle\pi\rangle\varphi \mid \langle\Lambda\rangle\varphi \end{aligned}$$

We finally expand the semantic definition for the transformation modalities:

$$[\![\langle\Lambda\rangle\varphi]\!]^M \quad = \quad [\![\varphi]\!]^{\Lambda(M)}$$

An accustomed reader or a keen logician might now expect me to axiomatise LPD and prove completeness. I will not do so in this paper. As is common in the dynamic epistemic logic literature, completeness for dynamics is not a difficult technical problem, because it can be avoided. In the case of LPD, we can use a translation $\varphi^\Lambda$ of formulas of the LPD language with transformation modalities to formulas without them, so that:

$$(w,a) \in [\![\varphi^\Lambda]\!]^M \quad \text{iff} \quad (w,a) \in [\![\varphi]\!]^{\Lambda(M)}$$

Whereas a transformation $\Lambda$ operates on a model $M$ to create a new model $\Lambda(M)$, the translation of formulas encodes the result of the transformation in the language without transformation modalities. It's as though the static language could predict what will happen after models are transformed.

The proof strategy I used is directly borrowed from the GDDL logic of Girard et al. (2012) and is straightforwardly adapted to the LPD context. The translation $\varphi^\Lambda$ needed for the reduction is the following:

$$\begin{array}{llll} p^\Lambda & = p & K^\Lambda & = \Lambda(K) \\ (\neg\varphi)^\Lambda & = \neg\varphi^\Lambda & F^\Lambda & = \Lambda(F) \\ (\varphi\wedge\psi)^\Lambda & = (\varphi^\Lambda\wedge\psi^\Lambda) & H_a^\Lambda & = \Lambda(H_a) \\ (\langle\pi\rangle\varphi)^\Lambda & = \langle\pi^\Lambda\rangle\varphi^\Lambda & (H_a^<)^\Lambda & = (H_a^\Lambda)^< \\ & & (\pi_1\cup\pi_2)^\Lambda & = \pi_1^\Lambda\cup\pi_2^\Lambda \\ & & (\pi_1;\pi_2)^\Lambda & = \pi_1^\Lambda;\pi_2^\Lambda \\ & & (\pi^*)^\Lambda & = (\pi^\Lambda)^* \\ & & (\varphi?)^\Lambda & = (\varphi^\Lambda)? \end{array}$$

With this translation, a straightforward induction establishes Lemma 1, which states that the logic with transformations can be systematically reduced to one without them:

[8] We only accept transformations that produce hierarchical models. Here's a technical problem for the inclined reader: how do you characterise acceptable transformations for different logics? That is, if I give you a class of models, how can you isolate transformations that will produce models within the same class?

**Lemma 1.** *For each world-agent pair* $(w, a)$ *of* $\Lambda(M)$ *and* $(v, b)$ *of* $M$*, and for each formula* $\varphi$*:*

$$\begin{array}{lll} (w,a) \in [\![\varphi^{\Lambda}]\!]^{M} & \textit{iff} & (w,a) \in [\![\varphi]\!]^{\Lambda(M)} \\ \langle (w,a),(v,b)\rangle \in [\![\pi^{\Lambda}]\!]^{M} & \textit{iff} & \langle (w,a),(v,b)\rangle \in [\![\pi]\!]^{\Lambda(M)} \end{array}$$

Therefore, as far as completeness of the logic is concerned, no additional work is required to axiomatise $\langle\Lambda\rangle$ modalities.

## 7 Conclusion

This concludes our investigation of promotion and demotion as can be expressed in LPD. Many more operations can be defined in LPD, but I have selected those which I think are most interesting. I have left some topics untouched in this paper. In particular, I haven't mentioned anything about the axiomatisation of the static part of LPD. Although not a trivial task, I believe this will not present serious difficulties. The axiomatisation of the hierarchical modalities would be based on that for total preorders for $[H_a]\varphi$ with an axiom for the proper interaction with $[H_a^<]$: $a \to (H_b^<\varphi \leftrightarrow H_b(\varphi \wedge \neg H_b a))$. One also needs an axiom for the proper interaction with the friendship modality: $[F]\varphi \to [H_a]\varphi$. Another aspect of the GDDL which I haven't exploited is the formalisation of *private* actions, in which agents secretly change the hierarchy of their friends with some of them being ignorant of the change.

I have only considered operations of promotion and demotion on groups as they were suggested by the doxastic operation of revision and contraction found in the AGM literature. But the LPD language is very flexible, and we could use it to formalise a range of different notions of promotion and demotion. One could for instance define operations in which a demotion of $\varphi$-agents would put all $\varphi$-agents under all $\neg\varphi$-agents, or would put all best $\varphi$-agents under all best $\neg\varphi$-agents; with neither of these alternative definitions would best $\varphi$-agents remain best friends after the demotion, as is the case when we use my own definitions. The preliminary framework and results I have provided encourage further investigation in a number of different directions.

## References

Andréka, Hajnal, Mark Ryan, and Pierre-Yves Schobbens. 2002. Operators and laws for combining preference relations. *Journal of Logic and Computation* 12(1): 13–53.

Girard, Patrick. 2011. Modal logic for lexicographic preference aggregation. In *Games, norms and reasons*, 97–117. Dordrecht: Springer.

Girard, Patrick, and Hans Rott. 2014. Belief revision and dynamic logic. In *Trends in logic, outstanding contributions: Johan F. A. K. Van Benthem on logical and informational dynamics*, vol. 5, ed. Alexandru Baltag and Sonja Smets. Cham: Springer.

Girard, Patrick, and Jeremy Seligman. 2009. An analytic logic of aggregation. In *Logic and its applications*, Lecture notes in computer science, vol. 5378, ed. R. Ramanujam and Sundar Sarukkai, 146–161. Berlin/Heidelberg: Springer.

Girard, Patrick, Jeremy Seligman, and Fenrong Liu. 2012. General dynamic dynamic logic. *Advances in Modal Logics* 9: 239–260

Harel, David, Jerzy Tiuryn, and Dexter Kozen. 2000. *Dynamic logic*. Cambridge: MIT.

Rott, Hans. 2009. Shifting priorities: Simple representations for twenty-seven iterated theory change operators. In *Towards mathematical philosophy*, Trends in logic, vol. 28, ed. David Makinson, Jacek Malinowski, and Heinrich Wansing, 269–296. Dordrecht: Springer.

Seligman, Jeremy, Fenrong Liu, and Patrick Girard. 2011. Logic in the community. In *Logic and Its Applications*, Lecture notes in computer science, vol. 6521, ed. Mohua Banerjee and Anil Seth, 178–188. Berlin/Heidelberg: Springer.

Seligman, Jeremy, Fenrong Liu, and Patrick Girard. 2013. Facebook and the epistemic logic of friendship. In *TARK XIV: Proceedings of the 13th conference on theoretical aspects of rationality and knowledge*, Chennai: India.

van Benthem, Johan. 2007. Dynamic logic for belief revision. *Journal of Applied Non-classical Logic* 17(2): 129–155.

# On the Attitude of Trust: A Formal Characterization of *Trust*, *Distrust*, and Associated Notions

**Andrew J.I. Jones**

**Abstract** Using modal logics to represent an agent's *beliefs*, *knowledge* and *wants*, an analysis is given of *trust* in terms of an agent's certainty that a particular, desired state-of-affairs will be realized. Similarly, a corresponding analysis of *distrust* is given. Placing these formal representations of *trust* and *distrust* at each of the ends of a spectrum, four intermediary structures may be identified, representing *hope*, two species of *anxiety*, and *fear*. In this way the relationships between the attitudes of trust/distrust and some basic types of emotional state may be precisely articulated. Some suggestions are also made regarding the analysis of some more complex types, including *regretting* that one trusted, and *being ashamed* that one trusted.

The paper employs modalities of type KD and KT for, respectively, the logics of *belief* and *knowledge*. It is shown that use of stronger doxastic and epistemic logics – of the kind often favoured in Artificial Intelligence – containing the positive and negative introspection axioms, would make three of the spectrum's four intermediary positions logically inconsistent. It is suggested that this result provides good reason for rejecting the stronger logics, and that their adoption in AI has often been motivated primarily by considerations of computational convenience, rather than by considerations of conceptual accuracy.

**Keywords** Trust • Distrust • Hope • Anxiety • Fear • Regret • Shame • Formal taxonomy of the emotions • Applied modal logic

Dedicated to the memory of my former mentor and research colleague Ingmar Pörn, who died in February 2014.

A.J.I. Jones (✉)
Department of Informatics, King's College London, London, UK

Imperial College London, London, UK
e-mail: andrewji.jones@kcl.ac.uk

A. Herzig, E. Lorini (eds.), *The Cognitive Foundations of Group Attitudes and Social Interaction*, Studies in the Philosophy of Sociality 5,
DOI 10.1007/978-3-319-21732-1_6

# 1 Introduction

Consider the following examples in which one agent trusts another:

- *x* trusts *y* to fulfil a contractual obligation;
- *x* trusts *y* to fulfil properly a role;
- *x* trusts what *y* says.

In Jones (2002) it was argued that, in each of these examples, the *content*, or *object*, of *x*'s trusting attitude concerns trustee compliance: *y*'s conformity to some governing norms or conventions. For the first example, the case is obvious; the contractual obligation is specified by some norm or other, and what *x* trusts is that *y* will comply with that norm. In the second example, the case turns on the assumption that roles are characterized, in part, in terms of a set of norms that apply to the role-holder when he is acting in that role – *cf.* (Pörn 1977, pp. 61–63). Trusting one's physician, for instance, amounts to trusting that he acts in ways that conform to the standards governing members of the medical profession. For the third example, the case turns on an assumption to the effect that indicative signalling, verbal or non-verbal, exploits conventions that correlate signalling act-types to types of states of affairs; when an instance of a given signalling act-type is performed, the conventionally correlated state of affairs *ought* then to hold. (For a detailed development of this approach to indicative signalling, see Jones and Kimbrough (2008); Jones and Parent (2007). The origin of the approach lies in Stenius (1967).)

In what follows, this 'trustee-compliance' view of the object of the trusting attitude will be presupposed.[1] The focus here will be not on *the object of* the trusting attitude, but rather on the *trusting attitude itself.* And in regard to the characterization of that attitude, Jones (2002) fell short in at least two respects:

- it described the *cognitive* aspect of the truster's attitude in terms of mere belief; but the fully trusting agent feels sure, certain, secure that trustee compliance will occur;
- it overlooked the *volitional* component.

The second of these two points reflects the fact that, ordinarily, it *matters* to truster *x* that compliance is forthcoming; compliance is not an issue on which *x* is indifferent: compliance is something that he wants. The presence of the volitional component in the trusting attitude explains, at least in part, why *trust* is so often linked to the notion of *risk*.

In order to develop an improved account of the trusting attitude, capable of repairing the shortcomings of the earlier approach, the point of departure here will be Pörn's modal-logical taxonomy of types of emotions (Pörn 1986), in which modal logics are used to represent the cognitive and volitional components alluded

[1] In my opinion most, if not all, other typical examples of situations in which one agent trusts another can also be understood in terms of this 'trustee-compliance' view of the object of the trusting attitude; but I shall not here argue that case.

to above.[2] One consequence will be that it becomes possible to get a clearer picture of the relationship between *trust* and *distrust*, on the one hand, and the cognitive and volitional aspects of *hope*, *anxiety* and *fear*, on the other.

## 2 Cognitive and Volitional Positions

Pörn (1986) applied the combinatory method of maxi-conjunctions, developed by Kanger for classifying types of Hohfeldian rights-relations.[3]

For the logic of *belief*, a modality of type KD is used, with the operator relativized to individual agents. The system KD of modal logic is formed by adding to the smallest normal system K – as defined in Chellas (1980) – the schema D:

D $B_x p \to \neg B_x \neg p$

which says that if an agent $x$ believes that $p$, where $p$ is any proposition, then he does not believe not-$p$. For the logic of *knowledge*, a modality of type KT is used, with the operator again relativized to individual agents. The system KT of modal logic is formed by adding to the smallest normal system K the schema T:

T $K_x p \to p$

A central conjecture in Pörn (1986) is that an agent's *certainty that p* may be represented as the agent's *believing that he knows that p*. Accordingly, the following two *certainty* positions may be identified:

$B_x K_x p$: $x$ is certain that $p$
$B_x K_x \neg p$: $x$ is certain that $\neg p$

In virtue of the logical properties of the two modalities $B$ and $K$, as modalities of type KD and KT, respectively, the following relations of logical implication may be shown to hold between the two certainty positions and other, weaker doxastic-epistemic positions:

$B_x K_x p \to \neg B_x \neg K_x p \to \neg B_x K_x \neg p$
$B_x K_x p \to B_x \neg K_x \neg p \to \neg B_x K_x \neg p$
$B_x K_x \neg p \to \neg B_x \neg K_x \neg p \to \neg B_x K_x p$
$B_x K_x \neg p \to B_x \neg K_x p \to \neg B_x K_x p$

The class of doxastic-epistemic 'positions' may now be generated as follows: first take the four positive expressions $B_x K_x p$, $B_x K_x \neg p$, $B_x \neg K_x p$, $B_x \neg K_x \neg p$, and then

[2] Not long after the publication of Jones (2002), Ingmar Pörn suggested to me in conversation that the account there put forward had completely overlooked the *affective* aspect of trust. The present paper aims, in part, to remedy that oversight.

[3] For an overview of the method of generating 'normative positions', and for references to the work of Kanger and Hohfeld, see Jones and Sergot (1993).

the corresponding negative expressions $\neg B_x K_x p$, $\neg B_x K_x \neg p$, $\neg B_x \neg K_x p$, $\neg B_x \neg K_x \neg p$. These eight expressions can be arranged as four truth-functional tautologies:

1. $B_x K_x p \; v \; \neg B_x K_x p$
2. $B_x K_x \neg p \; v \; \neg B_x K_x \neg p$
3. $B_x \neg K_x p \; v \; \neg B_x \neg K_x p$
4. $B_x \neg K_x \neg p \; v \; \neg B_x \neg K_x \neg p$

Obviously, for any given agent, and for any proposition $p$, precisely one of the disjuncts in each of (1)–(4) must hold. There are 16 ways of selecting precisely one disjunct from each of (1)–(4), to form 16 conjunctions of four conjuncts each. Of these 16 conjunctions, just 6 are logically consistent, given the logical properties adopted for the two modal operators. The 6 logically consistent conjunctions are:

(DE1) $B_x K_x p \;\&\; \neg B_x \neg K_x p \;\&\; \neg B_x K_x \neg p \;\&\; B_x \neg K_x \neg p$
(DE2) $\neg B_x K_x p \;\&\; B_x \neg K_x p \;\&\; B_x K_x \neg p \;\&\; \neg B_x \neg K_x \neg p$
(DE3) $\neg B_x K_x p \;\&\; B_x \neg K_x p \;\&\; \neg B_x K_x \neg p \;\&\; B_x \neg K_x \neg p$
(DE4) $\neg B_x K_x p \;\&\; B_x \neg K_x p \;\&\; \neg B_x K_x \neg p \;\&\; \neg B_x \neg K_x \neg p$
(DE5) $\neg B_x K_x p \;\&\; \neg B_x \neg K_x p \;\&\; \neg B_x K_x \neg p \;\&\; B_x \neg K_x \neg p$
(DE6) $\neg B_x K_x p \;\&\; \neg B_x \neg K_x p \;\&\; \neg B_x K_x \neg p \;\&\; \neg B_x \neg K_x \neg p$

It may be shown that these six positions are mutually exclusive, and their disjunction is a logical truth. So precisely one of (DE1)–(DE6) must hold for any given proposition $p$.

Concerning (DE6), Pörn said (Pörn 1986), p. 208 that it "...is the epistemic null-position; it is doubtful whether it is relevant for the analysis of emotions since in this position the subject has no belief at all concerning $p$. (An epistemic null-position may of course be the object of an emotion, but that is another matter.)" And in the development of his analysis of atomic emotional types he chose to disregard (DE6). In what follows, however, the possibility will be left open that (DE6) might be of relevance, particularly in the context of comparing *trust* with the (doxastic-epistemic components of) emotions. So (DE6) will be retained.

Each of the six (DE) positions may be simplified by removing any conjuncts that are themselves logically implied by one or more other conjunct. The result of that simplification is as follows:

(SDE1) $B_x K_x p$
(SDE2) $B_x K_x \neg p$
(SDE3) $B_x \neg K_x p \;\&\; B_x \neg K_x \neg p$
(SDE4) $B_x \neg K_x p \;\&\; \neg B_x K_x \neg p \;\&\; \neg B_x \neg K_x \neg p$
(SDE5) $B_x \neg K_x \neg p \;\&\; \neg B_x K_x p \;\&\; \neg B_x \neg K_x p$
(SDE6) $\neg B_x \neg K_x p \;\&\; \neg B_x \neg K_x \neg p$

As regards the logic of volition, let expressions of the form $D_x p$ be read 'x desires/wants that $p$', where $D_x$ is a (relativized) normal modality of type KD. It may then readily be shown that there are just three basic volitional positions for any given agent $x$ and any proposition $p$. They are:

(V1) $D_x p$
(V2) $D_x \neg p$
(V3) $\neg D_x p \mathbin{\&} \neg D_x \neg p$

Following Pörn, (V3) may be said to be the position of 'volitional indifference'. He was inclined to the view that it is irrelevant to the analysis of the emotions; given the present interest in the analysis of *trust*, and given what was said in the introductory remarks to the effect that it ordinarily *matters to* truster $x$ that trustee $y$ acts in a way that fulfils the trust bestowed upon him, the focus here will also be exclusively on volitional positions (V1) and (V2); indifference will be disregarded.

## 3 Cognitive and Volitional Positions Combined

The result of conjoining (V1) and (V2), respectively, to each of (SDE1)–(SDE6) is given in the following list of 12 doxastic-epistemic/volitional positions:

(DEV1) $B_x K_x p \mathbin{\&} D_x p$
(DEV2) $B_x K_x p \mathbin{\&} D_x \neg p$
(DEV3) $B_x K_x \neg p \mathbin{\&} D_x p$
(DEV4) $B_x K_x \neg p \mathbin{\&} D_x \neg p$
(DEV5) $B_x \neg K_x p \mathbin{\&} B_x \neg K_x \neg p \mathbin{\&} D_x p$
(DEV6) $B_x \neg K_x p \mathbin{\&} B_x \neg K_x \neg p \mathbin{\&} D_x \neg p$
(DEV7) $B_x \neg K_x p \mathbin{\&} \neg B_x K_x \neg p \mathbin{\&} \neg B_x \neg K_x \neg p \mathbin{\&} D_x p$
(DEV8) $B_x \neg K_x p \mathbin{\&} \neg B_x K_x \neg p \mathbin{\&} \neg B_x \neg K_x \neg p \mathbin{\&} D_x \neg p$
(DEV9) $B_x \neg K_x \neg p \mathbin{\&} \neg B_x K_x p \mathbin{\&} \neg B_x \neg K_x p \mathbin{\&} D_x p$
(DEV10) $B_x \neg K_x \neg p \mathbin{\&} \neg B_x K_x p \mathbin{\&} \neg B_x \neg K_x p \mathbin{\&} D_x \neg p$
(DEV11) $\neg B_x \neg K_x p \mathbin{\&} \neg B_x \neg K_x \neg p \mathbin{\&} D_x p$
(DEV12) $\neg B_x \neg K_x p \mathbin{\&} \neg B_x \neg K_x \neg p \mathbin{\&} D_x \neg p$

(DEV1)–(DEV10) are Pörn's ten 'atomic emotional types'. It is interesting to consider the labels he gave to them. (DEV1) and (DEV4) are both types of *security*, in the sense that, in each case, what the agent is certain of *matches* what he desires; by contrast, (DEV2) and (DEV3) represent *despair* (Pörn's label), or *hopelessness*, since in each case what the agent is certain of is the *opposite* of that which he desires.[4] (DEV5) and (DEV6) both represent a form of *anxiety*, in as much as the agent believes that he does not know whether $p$ holds – and in the one case he wants $p$, whereas in the other he doesn't. Consider next (DEV7): the agent desires $p$ and (first conjunct) believes that he does not know that $p$; although (second conjunct) he does not believe that he knows not-$p$, his knowing not-$p$ is compatible with all

[4] Consider the renowned cartoon-style picture by Roy Lichtenstein of the face of a young woman, resting on a pillow, tears flowing, thinking to herself 'That's the way it *should* have *begun*! But it's hopeless!' The situation is essentially that described by (DEV3): she is certain that it (the affair??) did *not* begin in that way, and she wishes that it *had*.

that he believes (third conjunct). (That description of the third conjunct follows (Hintikka 1962) by interpreting '$\neg B_x \neg$' as 'compatibility with all that $x$ believes'.) So (DEV7), and its counterpart (DEV10), represent *fear*. Parallel considerations lead to the conclusion that (DEV8) and (DEV9) represent *hope*.

(DEV11) and (DEV12) of course do not figure among Pörn's atomic types, since they are based on the epistemic null-position. But perhaps a case can be made for maintaining that they represent *another* type of *anxiety*. If it is compatible with all that an agent believes that he knows that $p$, but also compatible with all that he believes that he knows that not-$p$, then it would seem that he totally lacks any information that would enable him to decide the question of $p$'s truth/falsity. But then if he is not indifferent, either wanting $p$ to be the case or wanting not-$p$ to be the case, he has grounds for anxiety – albeit grounds of a cognitive type different from that expressed in (DEV5) and (DEV6). This point will be considered further below, in the discussion of *trust* and *distrust*.

It is important to note Pörn's emphasis that – as he put it – the atomic types are 'unrestricted', in as much as their characterization is independent of any particular specification of the kind of state-of-affairs $p$ describes. He then considers ((Pörn 1986), pp. 209–210), by way of contrast, some examples of emotions, such as *anger*, that *are* 'restricted to objects of a certain kind'. It is at this point that the above account of the doxastic-epistemic/volitional positions can be linked to the introductory discussion of the object of the trusting attitude.

## 4 A Spectrum of Cases

Assume now that the proposition $p$ in (DEV1)–(DEV12) is restricted to *trustee compliance* in the sense described in the first paragraph of this paper, and elaborated in (Jones 2002). It is appropriate then to confine attention to those six cases in which $D_x p$, rather than $D_x \neg p$, appears, since the assumption is that truster x desires compliance on the part of the trustee. The contracted list is this:

(DEV1) $B_x K_x p \ \& \ D_x p$
(DEV3) $B_x K_x \neg p \ \& \ D_x p$
(DEV5) $B_x \neg K_x p \ \& \ B_x \neg K_x \neg p \ \& \ D_x p$
(DEV7) $B_x \neg K_x p \ \& \ \neg B_x K_x \neg p \ \& \ \neg B_x \neg K_x \neg p \ \& \ D_x p$
(DEV9) $B_x \neg K_x \neg p \ \& \ \neg B_x K_x p \ \& \ \neg B_x \neg K_x p \ \& \ D_x p$
(DEV11) $\neg B_x \neg K_x p \ \& \ \neg B_x \neg K_x \neg p \ \& \ D_x p$

A suitable label for (DEV1) is *complete trust*: $x$ is certain that compliance, which he desires, will be forthcoming. Thus *complete trust* is a specific instance of what Pörn calls *security*. By contrast (DEV3) represents *complete distrust*: $x$ is certain that compliance, which he desires, will *not* be forthcoming. Thus, plausibly enough, *complete distrust* is a specific instance of *despair* or *hopelessness*.

The six cases might in fact be considered to constitute a spectrum, with *complete trust* at the left-hand end and *complete distrust* at the right-hand end. Next to *complete trust* comes (DEV9), which represents *hope-of-compliance*, and immediately preceding *complete distrust* comes (DEV7), which represents *fear-of-non-compliance*. The middle of the spectrum is occupied by the positions (DEV5) and (DEV11), which represent two species of *anxiety-about-whether-compliance-will-occur*. So the spectrum looks like this:

(DEV1) – (DEV9) – [(DEV5), (DEV11)] – (DEV7) – (DEV3)

One of the contexts in which matters of *trust* have lately been given a great deal of attention is the field of e-commerce. In that context, a distinction is often drawn between commercial interactions in which the traders have some previous experience of each other, and the so-called 'first-trade scenario', where the parties may be completely unfamiliar with one another. That distinction may perhaps be used to illustrate the differences between the two types of *anxiety* in the spectrum, (DEV5) and (DEV11). The latter fits, it seems, the kind of situation that would arise in a 'first-trade scenario' if the one party, $x$, totally lacks information relevant to assessing the trustworthiness of the other party, $y$, whereas (DEV5) would be a more appropriate description of the situation in which, on the basis of previous experience of $y$, $x$ has come to the conclusion that he just doesn't know whether or not $y$ can be trusted.

Another way of highlighting the difference between (DEV5) and (DEV11) is as follows: in virtue of its first conjunct, (DEV11) logically implies $\neg B_x \neg p$, and in virtue of its second conjunct it logically implies $\neg B_x p$. However, $B_x p$ may be consistently conjoined with (DEV5), and $B_x \neg p$ may be consistently conjoined with (DEV5) – but obviously not both, because of the D schema. (DEV11) is characterized by the agent's lack of relevant information; only when that lack is remedied can he move to a position that would be compatible either with the belief that $p$, or with the belief that not-$p$.

Some may object to the description of the spectrum offered above, and indeed more generally to Pörn's approach to the characterization of the emotions, on the grounds that there is more to an emotional state than the mere combination of epistemic-doxastic and volitional elements, making it inappropriate to use such terms as *hope*, *fear, anxiety* as labels. But nothing essential hinges on the use of those terms; the six positions, and their ordering on the spectrum, are clearly characterized by means of the component logics, and the entire account could thus be re-formulated without appeal to the emotion-terms. The key point is that a small set of modal building-blocks have been used to describe precisely and formally the attitudes of *complete trust* and *complete distrust*, and to exhibit their respective relationships to, and differences from, a set of intermediary attitudes.

## 5 Strengthening the Logics of Belief and Knowledge

It is commonly accepted that knowledge implies belief. So now add to the logics described above the schema:

KB $\quad K_x p \rightarrow B_x p$

Furthermore, it has been usual in Artificial Intelligence to adopt KD45 for the logic of belief and KT5 for the logic of knowledge. Essentially, this amount to adding to the logic KD (for belief), and the logic KT (for knowledge), the so-called *positive* and *negative introspection* schemas*:*

B4 $B_xp \rightarrow B_xB_xp$ *(positive introspection)*
B5 $\neg B_xp \rightarrow B_x\neg B_xp$ *(negative introspection)*
K4 $K_xp \rightarrow K_xK_xp$ *(positive introspection)*
K5 $\neg K_xp \rightarrow K_x\neg K_xp$ *(negative introspection)*

From the semantical point of view, this strengthened logic of knowledge and belief can be characterized by means of a standard model (in the sense of Chellas (1980)) in which there are two binary accessibility relations $R^K{}_x$ and $R^B{}_x$ satisfying the following properties:

$R^K{}_x$ is both reflexive and Euclidean
$R^B{}_x$ is serial, transitive and Euclidean
$R^B{}_x$ is a sub-relation of $R^K{}_x$.

The basic truth condition for sentences of the form $K_xp$ is given as follows:

(TCK) At any world $w$ in any standard model $M$, $K_xp$ is true at $w$ iff $p$ itself is true at every world $w_1$ such that $<w, w_1> \in R^K{}_x$

Similarly, the basic truth condition for sentences of the form $B_xp$ is given by:

(TCB) At any world $w$ in any standard model $M$, $B_xp$ is true at $w$ iff $p$ itself is true at every world $w_1$ such that $<w, w_1> \in R^B{}_x$.

Adoption of this strengthened logic of knowledge and belief would have significant consequences for the 'trust-distrust' spectrum described above, since *three* of the six component positions – (DEV7), (DEV9) and (DEV11) – would become *inconsistent*. In terms of the semantics, the inconsistencies may be demonstrated in the following way, starting from (DEV11).

Suppose that each of the first two conjuncts of (DEV11), $\neg B_x\neg K_xp$ and $\neg B_x\neg K_x\neg p$, is true at some world $w$ in a model $M$ of the kind just outlined above. Since $\neg B_x\neg K_xp$ holds at $w$, it follows by (TCB) that there must be some world $w_1$ such that $<w, w_1> \in R^B{}_x$ and such that $K_xp$ holds at $w_1$. Similarly, since $\neg B_x\neg K_x\neg p$ holds at $w$, it follows by (TCB) that there must be some world $w_2$ such that $<w, w_2> \in R^B{}_x$ and such that $K_x\neg p$ holds at $w_2$. Since $R^B{}_x$ is a sub-relation of $R^K{}_x$, it now follows that $<w, w_1> \in R^K{}_x$ and that $<w, w_2> \in R^K{}_x$. But the relation $R^K{}_x$ is Euclidean, so it now follows that $<w_1, w_2> \in R^K{}_x$. But then, since $K_xp$ holds at $w_1$, it follows by (TCK) that $p$ itself must hold at $w_2$. However, since $R^K{}_x$ is also reflexive, it follows that $<w_2, w_2> \in R^K{}_x$ and thus, since $K_x\neg p$ holds at $w_2$, it follows by (TCK) that $\neg p$ also holds at $w_2$. This reduces to absurdity the initial assumption that each of $\neg B_x\neg K_xp$ and $\neg B_x\neg K_x\neg p$ holds at $w$.

Consider next (DEV9), and suppose that its second and third conjuncts, $\neg B_xK_xp$ and $\neg B_x\neg K_xp$, are both true at some world $w$ in a model $M$ of the kind under consideration. Since $\neg B_x\neg K_xp$ holds at $w$, it follows by (TCB) that there must be some world $w_1$ such that $<w, w_1> \in \mathrm{R}^{\mathrm{B}}{}_{\mathrm{x}}$ and such that $K_xp$ holds at $w_1$. Similarly, since $\neg B_xK_xp$ holds at $w$, it follows by (TCB) that there must be some world $w_2$ such that $<w, w_2> \in \mathrm{R}^{\mathrm{B}}{}_{\mathrm{x}}$ and such that $\neg K_xp$ holds at $w_2$. Since $\mathrm{R}^{\mathrm{B}}{}_{\mathrm{x}}$ is a sub-relation of $\mathrm{R}^{\mathrm{K}}{}_{\mathrm{x}}$, it now follows that $<w, w_1> \in \mathrm{R}^{\mathrm{K}}{}_{\mathrm{x}}$ and that $<w, w_2> \in \mathrm{R}^{\mathrm{K}}{}_{\mathrm{x}}$. But the relation $\mathrm{R}^{\mathrm{K}}{}_{\mathrm{x}}$ is Euclidean, so it now follows that $<w_2, w_1> \in \mathrm{R}^{\mathrm{K}}{}_{\mathrm{x}}$. Now, since $\neg K_xp$ holds at $w_2$, it follows by (TCK) that there must be some world $w_3$ such that $<w_2, w_3> \in \mathrm{R}^{\mathrm{K}}{}_{\mathrm{x}}$ and such that $\neg p$ holds at $w_3$. Since it has been established that $<w_2, w_1> \in \mathrm{R}^{\mathrm{K}}{}_{\mathrm{x}}$ and that $<w_2, w_3> \in \mathrm{R}^{\mathrm{K}}{}_{\mathrm{x}}$, it now follows from a further application of the Euclidean property of $\mathrm{R}^{\mathrm{K}}{}_{\mathrm{x}}$ that $<w_1, w_3> \in \mathrm{R}^{\mathrm{K}}{}_{\mathrm{x}}$. But $K_xp$ holds at $w_1$; so it follows by (TCK) that $p$ itself must hold at $w_3$. This reduces to absurdity the initial assumption that each of $\neg B_xK_xp$ and $\neg B_x\neg K_xp$ holds at $w$. (By means of the same pattern of argument it may also be demonstrated that (DEV7) is inconsistent.)

Apart from the basic truth conditions for the knowledge and belief modalities, these proofs of inconsistency turn on three properties: that $\mathrm{R}^{\mathrm{K}}{}_{\mathrm{x}}$ is Euclidean; that $\mathrm{R}^{\mathrm{K}}{}_{\mathrm{x}}$ is reflexive; and that $\mathrm{R}^{\mathrm{B}}{}_{\mathrm{x}}$ is a sub-relation of $\mathrm{R}^{\mathrm{K}}{}_{\mathrm{x}}$. The second and third properties are unproblematic: reflexivity is the key to guaranteeing validity of the T schema, and thus that knowledge implies truth, and the sub-relation property guarantees the KB schema, and thus that knowledge implies belief. So the problem lies in the adoption of the Euclidean property for the epistemic accessibility relation – the very property that plays the key role in validating the positive and negative introspection schemas for the knowledge modality. Given the intuitive plausibility of (DEV7), (DEV9) and (DEV11) as representations of, respectively, the *fear-of-non-compliance* position, the *hope-of-compliance* position, and one of the *anxiety* positions, the conclusion to be drawn is that the Euclidean property, and the corresponding introspection schemas, should be rejected.[5]

Why was KT5 often the epistemic logic of choice in AI ? Part of the answer to that question, perhaps, lies in the fact that KT5 has properties that are attractive from the point of view of *computational* tractability. The practice of allowing *computational* considerations to play a significant role in determining choice of *conceptual* model has been quite widespread in AI and in the closely allied research field of Agents and Multi-agent Systems. The problematic nature of that practice is discussed in some detail in (Jones et al. 2013), which outlines an approach to the design of intelligent socio-technical systems in which conceptual and computational models are properly *integrated*.

[5] In some ongoing work on the application of the modal logic of belief to the characterization of *self-deception*, I have reached the same type of conclusions regarding KD45; that work identifies a class of intuitively plausible 'self-deception positions', each of which can be consistently represented if the belief logic is of type KD, and all of which become logically inconsistent if the logic used is KD45. See Jones (2013).

## 6 Non-atomic Types

Pörn (1986, pp. 210–213), discusses ways in which atomic types of emotions can be combined to form complex types. Pörn considers, for instance, *envy*; suppose

(i) $x$ is envious of $y$ because $y$ got the job.

Here, he suggests, we have a situation in which two instances of *despair* are combined, in that

(ii) $x$ is certain that $y$ got the job, but wishes that he ($y$) had not; and
(iii) $x$ is certain that he himself did not get the job, but wishes that he had.

So, where $p =$ '$y$ got the job' and $q =$ '$x$ got the job', the logical form of (i) becomes:

(iv) $B_xK_xp \mathbin{\&} D_x\neg p \mathbin{\&} B_xK_x\neg q \mathbin{\&} D_xq$

Another way in which complex emotional types can be formed, Pörn suggests, is when the object of an emotion is itself also an emotion. Consider this in relation to the case of *complete trust*, as analysed above[6]:

(v) $x$ is certain that $p$ (i.e., that $y$ will comply), and $x$ desires that $p$.

And now consider how to interpret

(vi) *x regrets* putting complete trust in $y$.

What, according to (vi), is $x$'s attitude (doxastic-epistemic and volitional) towards the object of his regret, as expressed by (v) ? A natural answer is that $x$ is certain that (v) and desires that it had not been the case that (v). Expressed formally:

(vii) $B_xK_x(B_xK_xp \mathbin{\&} D_xp) \mathbin{\&} D_x\neg(B_xK_xp \mathbin{\&} D_xp)$

Consider a specific example: $x$ trusted Nick Clegg and the Liberal Democrats at the 2010 UK General Election; $x$ was certain (believed that he knew) that they would deliver on their (manifold) promises, and desired that they should do so. $x$ now regrets trusting: he is quite sure that he had that trust, and he wishes that he hadn't !

What then would be the difference between *that* situation and one in which $x$ is *ashamed of* having trusted Clegg and the Liberal Democrats ? One suggestion would be that $x$'s shame combines his regret with a conviction that he *ought not* to have trusted in the first place: he should have known better, should have been able to see through the pretence . . . . . . If a suggestion of that sort is accepted, then the modal-logical language needs to be supplemented with an appropriate normative modality to represent 'ought'. When that component is supplied, the sentence

[6] Although the account that now follows bears some similarities to Pörn's, it differs substantially from it in points of detail. In particular, I do not make use of the notion of the *appropriateness* or *inappropriateness* of an emotion.

(viii) $x$ is ashamed of putting complete trust in $y$

may be rendered as

(ix) $B_xK_x(B_xK_xp \; \& \; D_xp) \; \& \; D_x\neg(B_xK_xp \; \& \; D_xp) \; \& \; B_xK_x \; Ought\neg(B_xK_xp \; \& \; D_xp)$

## 7 Concluding Remark

The paper has indicated a way of placing the concepts of *trust* and *distrust* in relation to a broader class of attitudes, in which doxastic-epistemic and volitional components are combined. It is evident that much work remains to be done on exploring the cognitive, volitional and perhaps normative aspects of the structure of complex types of emotions. Hopefully, however, the discussion presented here provides grounds for thinking that, in the spirit of Pörn's work, these phenomena are amenable to rigorous and systematic analysis by means of application of the tools of modal logic.

**Acknowledgements** 1. A second paper by Ingmar Pörn on the emotions, written in Swedish under the title 'Jagvärderande emotioner' ('Emotions of self-assessment'), also proved useful in writing the present piece. I am most grateful to my daughter, Linnea Erasmie-Jones, for producing an English translation of it. The Swedish copy that I have – a photocopied typescript – is dated 1994. Although I do not have specific details, it was apparently intended for a collection of papers on the emotions to be published by a philosophical society in Sweden or Finland. 2. My thanks to Robert Demolombe for helpful comments on an earlier version of this paper.

## References

Chellas, Brian F. 1980. *Modal logic – An introduction*. Cambridge: Cambridge University Press.

Hintikka, Jaakko. 1962. *Knowledge and belief – An introduction to the logic of the two notions*. Ithaca/London: Cornell University Press.

Jones, Andrew J.I. 2002. On the concept of trust. In *Formal modeling and electronic commerce* (Special issue of Decision Support Systems, vol. 33, no. 3), ed. Steven Kimbrough et al., 225–232.

Jones, Andrew J.I. 2013. *On the formal-logical characterisation of self-deception*. Invited talk at ArgMAS 2013, a workshop at AAMAS 2013, St. Paul, Minnesota, May 2013. Draft paper nearing completion.

Jones, Andrew J.I., and Steven O. Kimbrough. 2008. The normative aspect of signalling and the distinction between performative and constative. *Journal of Applied Logic* 6(2): 218–228.

Jones, Andrew J.I., and Xavier Parent. 2007. A convention-based approach to agent communication languages. *Group Decision and Negotiation* 16: 101–141.

Jones, Andrew J.I., and Marek J. Sergot. 1993. On the characterisation of law and computer systems: The normative systems perspective, chapter 12. In *Deontic logic in computer science: Normative system specification* (Wiley professional computing series), ed. J.-J.Ch. Meyer, and R.J. Wieringa, 275–307. Wiley.

Jones, Andrew J.I., Alexander Artikis, and Jeremy V. Pitt. 2013. The design of intelligent socio-technical systems. *Artificial Intelligence Review* 39(1): 5–20.

Pörn, Ingmar. 1977. *Action theory and social science – Some formal models*, Synthese library, vol. 120. Dordrecht/Holland: Reidel.

Pörn, Ingmar. 1986. On the nature of emotions. In *Changing positions*, Philosophical Studies published by the University of Uppsala, Sweden: vol. 38, 205–214.

Stenius, Erik. 1967. Mood and language game. *Synthese* 17: 254–274.

# The Topology of Common Belief

**David Pearce and Levan Uridia**

**Abstract** We study a modal logic $\mathbf{K4}_{\mathbf{2}}^{\mathbf{C}}$ of common belief for normal agents. We discuss Kripke completeness and show that the logic has tree model property. A main result is to prove that $\mathbf{K4}_{\mathbf{2}}^{\mathbf{C}}$ is the modal logic of all $T_D$-intersection closed, bi-topological spaces with derived set interpretation of modalities. Based on the splitting translation we also discuss connections with $\mathbf{S4}_{\mathbf{2}}^{\mathbf{C}}$, the logic of common knowledge.

**Keywords** Modal logic • Epistemic logic • Common belief • Common knowledge

## 1 Introduction

In logics for knowledge representation and reasoning, the study of epistemic and doxastic properties of agents with certain, intuitively acceptable, restrictions on their knowledge and belief is a well-developed area. Smullyan (1986) discusses various types of agents based on properties of belief. In his terminology, an agent whose belief satisfies the modal axiom (4) : $\Box p \rightarrow \Box\Box p$, translated as 'If the agent believes $p$, then he believes that he believes $p$', is called a *normal agent*. **K4** is the modal logic which formalizes the belief behavior of normal agents. This generalizes the classical doxastic system **KD45** in the same way as **S4** generalizes the epistemic logic **S5**, by dropping some restrictions on the properties of an agent.

The study of *group attitudes* is already well-established in several fields where collective opinion and reasoning are important. Also in newly emerging areas such as agreement technologies, and 'social intelligence', iterative concepts of agent belief and knowledge are of special interest. To achieve successful communication and agreement it is important for agents to reason about themselves and what others

D. Pearce (✉)
Universidad Politécnica de Madrid, Madrid, Spain
e-mail: pearcedav@gmail.com

L. Uridia
TSU Razmadze Mathematical Institute, Tbilisi, Georgia

A. Herzig, E. Lorini (eds.), *The Cognitive Foundations of Group Attitudes and Social Interaction*, Studies in the Philosophy of Sociality 5,
DOI 10.1007/978-3-319-21732-1_7

know or believe. Among the more fundamental concepts are the notions of *common knowledge* and *common belief*. We denote the operators for common knowledge and common belief by $C_K$ and $C_B$ respectively. We have: $C_K\varphi$ iff $\varphi$ is common knowledge in the group $K$ and $C_B\varphi$ iff $\varphi$ is a common belief in the group $B$.

Following the analysis of common knowledge as originally defined by Lewis (1969), this concept has been extensively studied from various perspectives in philosophy (Barwise 1988; Aumann 1976), game theory (van Benthem 2007), artificial intelligence (Herzig et al. 2009), modal logic (Baltag et al. 1998; Baltag and Smets 2009; Bezhanishvili and van der Hoek 2014) etc. Theories of common belief are less well-developed though some approaches can be found in Stalnaker (2001), Herzig et al. (2009), and Lismont and Mongin (1994). The present chapter is devoted to a study of the common belief of 'normal' agents in the sense mentioned above. We want to extend and bring together two previous lines of work. One direction is our own study of several extensions of the modal logic **wK4** that form interesting doxastic logics different from **KD45**; see in particular Pearce and Uridia (2010, 2011a,b). **wK4** is the normal modal logic based on the axioms

$(K)$: $\Box(p \to q) \to \Box p \to \Box q$
$(w4)$: $\Box p \wedge p \to \Box\Box p$

In previous work we showed that different extensions of **wK4** may be useful in certain doxastic contexts, for instance in modeling a notion of *minimal* belief, and more generally for *non-monotonic* reasoning about beliefs. They provide alternatives to the more familiar system **KD45** and its non-monotonic extension, *autoepistemic* logic. We also considered topological interpretations and embedding relations between epistemic and doxastic logics, i.e. translations between knowledge and belief operators. However in our earlier studies only single agent systems are treated. In our view, these different extensions of **wK4** can all be considered types of *doxastic* logics, even if they omit or weaken some of the stronger epistemic axioms.[1]

Our second point of departure is provided by the work of van Benthem and Sarenac (2004), who showed how a topological semantics for logics of common knowledge may be useful for modeling and distinguishing different concepts. A key idea here is that the knowledge of different agents is represented by different topologies over a set $X$. Various ways to merge that knowledge can be obtained via different modes of combining logics and topological models. van Benthem and Sarenac (2004) considers for example the fusion logic $\mathbf{S4}\circ\mathbf{S4}$ and product topologies that are complete for the common knowledge logic $\mathbf{S4}_2^{\mathbf{C}}$ of Fagin et al. (1995).

In light of Fagin et al. (1995) and van Benthem and Sarenac (2004) and our previous work several natural questions emerge that we want to address. In summary the main tasks of this chapter are:

[1] Lismont and Mongin (1994), treating common belief, and Steinsvold (2008), treating topological models for belief, are related works that also study weaker extensions of **K4**.

1. Define a logic $\mathbf{K4}_\mathbf{2}^\mathbf{C}$ of common belief for normal agents and prove its completeness for a Kripke, relational semantics. Show it has the finite model property and the tree model property.
2. Study a topological semantics for $\mathbf{K4}_\mathbf{2}^\mathbf{C}$ and prove completeness for intersection topologies. Specifically show that $\mathbf{K4}_\mathbf{2}^\mathbf{C}$ is the modal logic of all $T_D$-intersection closed, bi-topological spaces with a derived set interpretation of modalities.
3. Belief under the topological interpretation of $\mathbf{K4}_\mathbf{2}^\mathbf{C}$ is understood via colimits and common belief in terms of colimits in the intersection topology. From 2 we aim to derive a topological condition for common belief in terms of colimits that is very similar to the corresponding condition that defines common knowledge in the modal $\mu$-calculus and is discussed at some length in van Benthem and Sarenac (2004).
4. Show how the common knowledge logic $\mathbf{S4}_\mathbf{2}^\mathbf{C}$ can be embedded in $\mathbf{K4}_\mathbf{2}^\mathbf{C}$ via the splitting translation that maps $C_K p$ into $p \wedge C_B p$.

## 1.1 Common Belief and the Topological Interpretation

As stated, we focus on the common belief of normal agents, and for ease of exposition we restrict ourselves to the two agent case. We thus consider two agents whose individual beliefs satisfy the axioms of **K4**. In other respects we adopt the main principles of the logic of common knowledge, $\mathbf{S4}_\mathbf{2}^\mathbf{C}$. This can be seen as a formalization of the idea that common knowledge is equivalent to an infinite conjunction of iterated individual knowledge: $\varphi \wedge \Box_1\varphi \wedge \Box_2\varphi \wedge \Box_1\Box_1\varphi \wedge \Box_1\Box_2\varphi \wedge \Box_2\Box_1\varphi \wedge \Box_2\Box_2\varphi \wedge \Box_1\Box_1\Box_1\varphi \wedge \Box_1\Box_1\Box_2\varphi \ldots$. Later we shall see that a variation of this formula is 'true' for common belief under the relational semantics. We shall also show that the topological semantics for $\mathbf{K4}_\mathbf{2}^\mathbf{C}$ is compatible with the idea of common belief as a fixpoint *equilibrium*, a notion used by Barwise (1988) to describe common knowledge that can be captured by an expression of the modal $\mu$-calculus.

Our approach to providing a topological semantics follows the work of Esakia (2001). Notice that under the topological interpretation of $\Box$ as a knowledge operator, e.g. in van Benthem and Sarenac (2004), $\Box\varphi$ refers to the topological *interior* of the points assigned to $\varphi$. In the case of a doxastic logic like **K4** our topological interpretation is different. It is perhaps simpler to state it for the $\Diamond$ operator. Following McKinsey and Tarski (1944), the idea is to treat $\Diamond\varphi$ as the *derivative* of the set $\varphi$ in the topological space. Esakia showed that under this interpretation **wK4** is the modal logic of all topological spaces. **K4** is an extension of **wK4** and is characterized in this semantics by the class of all $T_D$-spaces (Bezhanishvil et al. 2005). Steinsvold was one of the first to look at derived set semantics from a doxastic point of view (Steinsvold 2008, 2009). By combining the ideas and results from van Benthem and Sarenac (2004), Steinsvold (2009) and Esakia (2001), we can obtain a derived set semantics for the logic of common belief based on bi-topological spaces, where the modality for common belief operates on the intersection of the two topologies. As a main result, we can prove that $\mathbf{K4_2^C}$ is sound and complete with respect to the special subclass of all bi-topological $T_D$-spaces.

## 2 Logic of Common Belief

We turn to the syntax and Kripke semantics of the logic $\mathbf{K4_2^C}$. The interpretation of common belief operator $C_B$ on bi-relational Kripke frames is similar to the interpretation of the common knowledge operator $C_K$, and is based on the notion of transitive closure of a relation. In this section we show that the logic $\mathbf{K4_2^C}$ is sound and complete with respect to the class of all bi-relational transitive Kripke structures. The proof is a slight modification of the completeness proof for the logic $\mathbf{S4_2^C}$ given in Fagin et al. (1995) therefore we only sketch the essential parts where the difference shows up. Additionally we show that every non-theorem of $\mathbf{K4_2^C}$ can be falsified on an infinite, irreflexive, bi-transitive tree.

### 2.1 Iterative Common Belief

There are different notions of common belief (Barwise 1988). Let us mention common belief as an infinite conjunction of nested beliefs and common belief as an equilibrium. Under the former idea, a proposition $p$ is a common belief of two agents if: agent-1 believes that $p$ and agent-2 believes that $p$ and agent-1 believes that agent-2 believes that $p$ and agent-2 believes that agent-1 believes that $p$ etc., where all possible finite mixtures occur. If we formalize this idea in a modal language with belief operators $\Box_1$ and $\Box_2$ for each agent respectively, then we arrive at the following concept of a common belief operator $C_B^\omega$.

$$C_B^0 p = \Box_1 p \wedge \Box_2 p;$$

$$C_B^{n+1} p = \Box_1 C_B^n p \wedge \Box_2 C_B^n p;$$

$$C_B^\omega p = \bigwedge\nolimits_{n\in\omega} C_B^n p.$$

$C_B^\omega$ exactly formalizes the intuition behind the former idea of common belief. However, since $C_B^\omega$ is an infinite intersection, it cannot be expressed as an ordinary formula of modal logic and hence studied in the usual approaches to standard modal logic. Nevertheless it turns out that we can capture the infinitary behavior of $C_B^\omega$ in a finitary sense. This idea is made more precise via the modal logic $\mathbf{K4_2^C}$.

## 2.2 Syntax

Throughout we work in the modal language $\mathcal{L}_C$ with an infinite set *Prop* of propositional letters and symbols $\wedge, \neg, \Box_1, \Box_2, C_B$. The set of formulas *Form* is constructed in a standard way: $Prop \subseteq Form$. If $\alpha, \beta \in Form$ then $\neg\alpha, \alpha \wedge \beta, \Box_1\alpha, \Box_2\alpha, C_B\alpha \in Form$. We will use standard abbreviations for disjunction and implication, $\alpha \vee \beta \equiv \neg(\neg\alpha \wedge \neg\beta)$ and $\alpha \to \beta \equiv \neg\alpha \vee \beta$.

- The axioms of the logic $\mathbf{K4^C_2}$ are all classical tautologies, each box satisfies all **K4** axioms, i.e. we have:

$$(K): \ \Box_i(p \to q) \to (\Box_i p \to \Box_i q)$$
$$(4): \Box_i p \to \Box_i\Box_i p$$

  for each $i \in \{1,2\}$ and in addition we have the equilibrium axiom for the common belief operator:

$$(equi): C_B p \leftrightarrow \Box_1 p \wedge \Box_2 p \wedge \Box_1 C_B p \wedge \Box_2 C_B p.$$

- The rules of inference are: Modus-Ponens, Substitution, Necessitation for $\Box_1$ and $\Box_2$ and the induction rule for the common belief operator:

$$(ind): \frac{\vdash \varphi \to \Box_1(\varphi \wedge \psi) \wedge \Box_2(\varphi \wedge \psi)}{\vdash \varphi \to C_B\psi}$$

  where $\varphi$ and $\psi$ are arbitrary formulas of the language.

## 2.3 Kripke Semantics

The Kripke semantics for the modal logic $\mathbf{K4^C_2}$ is provided by transitive, bi-relational Kripke frames. The triple $(W, R_1, R_2)$, with $W$ an arbitrary set and $R_i \subseteq W \times W$ where $i \in \{1,2\}$, is a *bi-transitive Kripke frame* if both $R_1$ and $R_2$ are transitive relations. A quadruple $(W, R_1, R_2, V)$ is a bi-transitive Kripke model if $(W, R_1, R_2)$ is a bi-transitive Kripke frame and $V : Prop \to P(W)$ is a valuation function. Observe that we only have two relations, which give a semantics for $\Box_1$ and $\Box_2$. To interpret the common belief operator, $C_B$, we construct a new relation, which is a transitive closure of the union of $R_1$ and $R_2$.

**Definition 1.** The transitive closure $R^+$ of a relation $R$ is defined as the least transitive relation containing the relation $R$.

Two points $x$ and $y$ are related by the transitive closure of the relation if there exists a finite path $\langle x_1, \ldots, x_n\rangle$ starting at $x$ and ending at $y$.

**Definition 2.** For a given bi-relational Kripke model $\mathcal{M} = (W, R_1, R_2, V)$ the satisfaction of a formula at a point $w \in W$ is defined inductively as follows:

$w \Vdash p$ iff $w \in V(p)$,
$w \Vdash \alpha \wedge \beta$ iff $w \Vdash \alpha$ and $w \Vdash \beta$,
$w \Vdash \neg\alpha$ iff $w \nVdash \alpha$,
$w \Vdash \Box_i\varphi$ iff $(\forall v)(wR_iv \Rightarrow v \Vdash \varphi)$,
$w \Vdash C_B\varphi$ iff $(\forall v)(w(R_1 \cup R_2)^+v \Rightarrow v \Vdash \varphi)$.

A formula $\alpha$ is valid in a model $\mathcal{M}$, in symbols $\mathcal{M} \Vdash \alpha$, if for every point $w \in W$ we have $w \Vdash \alpha$. $\alpha$ is valid in a bi-relational frame $\mathcal{F} = (W, R_1, R_2)$, in symbols $\mathcal{F} \Vdash \alpha$, iff $\alpha$ is valid in every model $\mathcal{M} = (\mathcal{F}, V)$ based on the frame. $\alpha$ is valid in a class of bi-relational frames $K$ if for every frame $\mathcal{F} \in K$ we have $\mathcal{F} \Vdash \alpha$.

## 2.4 Soundness and Completeness

**Proposition 1 (Soundness).** *Modal logic* $\mathbf{K4_2^C}$ *is sound with respect to the class of all bi-transitive Kripke frames.*

*Proof.* The only non-trivial cases are to show that the equilibrium axiom and the induction rule hold in the class of all bi-transitive models. Let $\mathcal{M} = (W, R_1, R_2, V)$ be an arbitrary bi-transitive Kripke model. And let $w \in W$. Assume $w \Vdash C_B\varphi$. Let us first show that $w \Vdash \Box_1\varphi$. Take an arbitrary $v \in W$ such that $wR_1v$. This implies that $w(R_1 \cup R_2)^+v$ hence $v \Vdash \varphi$. Let us show that $w \Vdash \Box_1 C_B\varphi$. Take an arbitrary $v$ and $v'$ such that $wR_1v$ and $v(R_1 \cup R_2)^+v'$. By Definition 1 this means that there exists a finite path $\langle v_1, \ldots, v_n\rangle$ such that each $v_i(R_1 \cup R_2)v_{i+1}$ and $v_1 = v$ and $v_n = v'$. Then the new path $\langle w, v_1, \ldots, v_n\rangle$ is also finite going from $w$ to $v'$. Hence $w(R_1 \cup R_2)^+v$ which implies that $v \Vdash \varphi$. In the same way we prove that $w \Vdash \Box_2\varphi \wedge \Box_2 C_B\varphi$.

For the other direction assume $w \nVdash C_B\varphi$. By Definition 2 this means that there is a finite path $\langle v_1, \ldots, v_n\rangle$ such that each $v_i(R_1 \cup R_2)v_{i+1}$ and $v_1 = w$ and $v_n \nVdash \varphi$. Without loss of generality we can assume that $v_1R_1v_2$. In case $n = 2$ we have that $w \nVdash \Box_1\varphi$. In case $n > 2$ we have that $v_2 \nVdash C_B\varphi$, hence $w \nVdash \Box_1 C_B\varphi$.

Now let us show that the induction rule preserves the validity of formulas in a model. We show this by contraposition. Assume for some $\mathcal{M} = (W, R_1, R_2, V)$ we have $\mathcal{M} \nVdash p \to C_Bq$. This means that there is a point $w \in W$ with $w \Vdash p$ and $w \nVdash C_Bq$. This implies that there is a finite path $\langle w, v_1, \ldots, v_n\rangle$ starting from $w$ with $v_n \nVdash q$. Now we look at $v_{n-1}$. As far as $v_{n-1}(R_1 \cup R_2)v_n$ we have that $v_{n-1} \nVdash \Box_1(p \wedge q) \wedge \Box_2(p \wedge q)$. Now in case $v_{n-1} \Vdash p$ we get that $v_{n-1} \nVdash p \to \Box_1(p \wedge q) \wedge \Box_2(p \wedge q)$ hence $\mathcal{M} \nVdash p \to \Box_1(p \wedge q) \wedge \Box_2(p \wedge q)$. In case $v_{n-1} \nVdash p$ we repeat the procedure and move to $v_{n-2}$. By repeating this $n - 1$ times at most, either we find the point which falsifies $p \to \Box_1(p \wedge q) \wedge \Box_2(p \wedge q)$ or obtain that $v_1 \nVdash p$. The latter implies that $w \nVdash p \to \Box_1(p \wedge q) \wedge \Box_2(p \wedge q)$.

Before starting the completeness proof we introduce the special closure of a set of subformulas of a given formula. This set will serve as the carrier set for the Kripke model we construct to falsify a formula which is not a theorem of $\mathbf{K4_2^C}$. Assume $\varphi$ is an arbitrary formula. Let $Sub(\varphi)$ be the set of all sub-formulas of $\varphi$. Let $Sub^+(\varphi)$ denote the closure of $Sub(\varphi)$ in the following way: if $C_B\alpha \in Sub^+(\varphi)$ then the formulas $\Box_1\alpha$, $\Box_2\alpha$, $\Box_1 C_B\alpha$ and $\Box_2 C_B\alpha$ are also in $Sub^+(\varphi)$. Let $\sim Sub^+(\varphi)$ denote the closure of $Sub^+(\varphi)$ under a single negation. For readability reasons let us denote this set by $FL(\varphi)$ (another motivation for $FL(\varphi)$ is that this construction is very much alike to the Fisher-Ladner closure used in completeness proofs for propositional dynamic logic PDL (Fischer and Ladner 1979)).

**Proposition 2 (Completeness).** *Modal logic* $\mathbf{K4_2^C}$ *is complete with respect to the class of all finite, bi-transitive Kripke frames.*

*Proof.* Assume $\mathbf{K4_2^C} \nvdash \varphi$. Let $W$ be the set of all maximally consistent subsets of $FL(\varphi)$. Let us define the relations $R_1$ and $R_2$ on $W$ in the following way: For every $\Gamma, \Gamma' \in W$ we define $\Gamma R_x \Gamma'$ iff $(\forall\alpha)(\Box_x\alpha \in \Gamma \Rightarrow \Gamma' \vdash \alpha \wedge \Box_x\alpha)$, where $x \in \{1, 2\}$.

*Claim.* Each $R_x$ is transitive.

*Proof.* Assume $\Gamma R_x \Gamma' \wedge \Gamma' R_x \Gamma''$ and $\Box_x\alpha \in \Gamma$. This implies that both $\Box_x\alpha$ and $\alpha$ are in $FL(\varphi)$. By the definition of $\Gamma R_x \Gamma'$ we have $\Gamma' \vdash \alpha \wedge \Box_x\alpha$, which implies $\Gamma' \vdash \Box_x\alpha$. As $\Box_x\alpha \in FL(\varphi)$ and $\Gamma'$ is maximally consistent set, we get $\Box_x\alpha \in \Gamma'$. Now we use again the definition of $\Gamma' R_x \Gamma''$ and we get that $\Gamma'' \vdash \alpha \wedge \Box_x\alpha$. Hence $\Gamma R_x \Gamma''$

So far we have defined a finite set $W$ with two transitive relations $R_1, R_2$ on it. Let $R_{1\vee 2}$ denote the transitive closure of the union of relations $R_1$ and $R_2$ i.e., $R_{1\vee 2} = (R_1 \cup R_2)^+$. At this point we are able to prove the truth lemma with respect to the model $\mathcal{M} = (W, R_1, R_2, R_{1\vee 2}, V)$, where $\Gamma \Vdash p$ iff $p \in \Gamma$. The proof goes by analogy to the proof for the common knowledge operator given in Fagin et al. (1995).

**Lemma 1 (Truth).** *For every formula* $\alpha \in FL(\varphi)$ *and every point* $\Gamma \in W$ *of the model* $\mathcal{M}$*, the following equivalence holds:* $\Gamma \Vdash \alpha$ *iff* $\alpha \in \Gamma$.

The proof is by induction on the length of formula. The base case follows immediately from the definition of valuation. Assume for all $\alpha \in FL(\varphi)$ with length less then $k$ that: $\Gamma \Vdash \alpha$ iff $\alpha \in \Gamma$.

Let us prove the claim for $\alpha \in FL(\varphi)$ with length equal to $k$. If $\alpha$ is a conjunction or negation of two formulas then the result easily follows from the definition of the satisfaction relation and the properties of maximal consistent sets, so we can skip the proofs. Assume $\alpha = \Box_1\beta$ and assume $\Gamma \Vdash \alpha$. Take a set $B = \{\gamma : \Box_1\gamma \in \Gamma\} \cup \{\Box_1\gamma : \Box_1\gamma \in \Gamma\} \cup \{\neg\beta\}$. The sub-claim is that $B$ is inconsistent. Assume not, then there exists $\Gamma' \in W$ such that $\Gamma' \supseteq B$. This by definition of the relation $R_a$ means that $\Gamma R_1 \Gamma'$. This is because for every $\alpha$ if $\Box_1\alpha \in \Gamma$ then $\Gamma' \vdash \alpha$ and

$\Gamma' \vdash \Box_1\alpha$ and hence $\Gamma' \vdash \alpha \wedge \Box_1\alpha$. Now as $\neg\beta \in \Gamma'$, by inductive assumption we get $\Gamma' \Vdash \neg\beta$. Hence we get a contradiction with our assumption that $\Gamma \Vdash \Box_1\beta$. So $B$ is inconsistent. This means that there exists $\gamma_{i_1}, \gamma_{i_2}, \ldots \gamma_{i_n}, \Box_1\gamma_{j_1}, \ldots \Box_1\gamma_{j_m} \in B$ such that $\vdash \gamma_{i_1} \wedge \gamma_{i_2} \wedge \ldots \wedge \gamma_{i_n} \wedge \Box_1\gamma_{j_1} \wedge \ldots \wedge \Box_1\gamma_{j_m} \to \beta$. Now we take the bigger conjunct, in particular we add $\Box\gamma_i$ for every $\gamma_i$ occurring in the conjunction, so we get: $\vdash (\gamma_{i_1} \wedge \Box_1\gamma_{i_1}) \wedge (\gamma_{i_2} \wedge \Box_1\gamma_{i_2}) \wedge \ldots \wedge (\gamma_{i_n} \wedge \Box_1\gamma_{i_n}) \wedge \Box_1\gamma_{j_1} \wedge \ldots \wedge \Box_1\gamma_{j_m} \to \beta$. Applying the necessitation rule for $\Box_1$ and using axiom 4 we get $\vdash \Box_1\gamma_{i_1} \wedge \ldots \wedge \Box_1\gamma_{i_n} \wedge \Box_1\gamma_{j_1} \wedge \ldots \wedge \Box_1\gamma_{j_m} \to \Box_1\beta$ so $\Gamma \vdash \Box_1\beta$, hence as $\Box_1\beta \in FL(\varphi)$ we conclude that $\Box_1\beta \in \Gamma$.

We just showed the left-to-right direction of our claim for $\alpha = \Box_1\beta$. For the right-to-left implication assume $\Box_1\beta \in \Gamma$. By the definition of $R_1$ for every $\Gamma'$ with $\Gamma R_1 \Gamma'$ we have $\Gamma' \vdash \beta \wedge \Box_1\beta$. From this it follows that $\Gamma' \vdash \beta$. As $\beta \in FL(\varphi)$ it follows that $\beta \in \Gamma'$ so by the inductive assumption $\Gamma' \Vdash \beta$.

The most important case is when $\alpha$ is of the form $C_B\beta$. Assume $\Gamma \Vdash \alpha$. Let $D = \{\Gamma \in W : \Gamma \Vdash C_B\beta\}$ and let $\delta = \bigvee_{\Gamma \in D} \hat{\Gamma}$, where $\hat{\Gamma}$ is the conjunction of all formulas inside $\Gamma$. Observe that as $W$ is finite $\hat{\Gamma}$ is a formula in our language. We want to show that $\vdash \delta \to \Box_1(\delta \wedge \beta) \wedge \Box_2(\delta \wedge \beta)$. We do it piece by piece.

First we show $\vdash \delta \to \Box_1\beta$. This follows by an analogous argument to the previous claim. So let us take $B = \{\gamma : \Box_1\gamma \in \Gamma\} \cup \{\Box_1\gamma : \Box_1\gamma \in \Gamma\} \cup \{\neg\beta\}$. This set is inconsistent, otherwise there would exist $\Gamma' \in W$ with $\Gamma R_1$ *Gamma*$'$ and $\Gamma' \not\Vdash \beta$, which contradicts $\Gamma \Vdash C_B\beta$. From the inconsistency of $B$ by the same argument as in the first claim it follows that $\vdash \hat{\Gamma} \to \Box_1\beta$. As $\Gamma$ was chosen arbitrarily we have $\vdash \delta \to \Box_1\beta$. Analogously we obtain $\vdash \delta \to \Box_2\beta$.

Now let us show that $\delta \to \Box_1\delta$. For this we take an arbitrary $\Gamma \in D$ and arbitrary $\Gamma' \notin D$ and show $\vdash \hat{\Gamma} \to \Box_1\neg\hat{\Gamma}'$. As $\Gamma \in D$, we have that $\Gamma \Vdash C_B\beta$, while for $\Gamma'$ we have $\Gamma' \not\Vdash C_B\beta$. This implies that not $\Gamma R_1 \Gamma'$, so by the definition of $R_1$, there is a formula $\psi$, such that $\Box_1\psi \in \Gamma$, while $\Gamma' \nvdash \Box_1\psi \wedge \psi$. From $\Gamma' \nvdash \Box_1\psi \wedge \psi$ we conclude that $\Box_1\psi \notin \Gamma'$ or $\psi \notin \Gamma'$. Now as both $\psi$ and $\Box_1\psi$ are in $FL(\varphi)$ we have $\neg\Box_1\psi \in \Gamma'$ or $\neg\psi \in \Gamma'$. This means that $\hat{\Gamma}'$ has the form either $\neg\Box_1\psi \wedge \psi \wedge \bigwedge \gamma_i$ or $\neg\Box_1\psi \wedge \neg\psi \wedge \bigwedge \gamma_i$ or $\Box_1\psi \wedge \neg\psi \wedge \bigwedge \gamma_i$. Then $\neg\hat{\Gamma}'$ is of the form $\Box_1\psi \vee \neg\psi \vee \bigvee \neg\gamma_i$ or $\Box_1\psi \vee \psi \vee \bigvee \neg\gamma_i$ or $\neg\Box_1\psi \vee \psi \vee \bigvee \neg\gamma_i$. In each case $\vdash \Box_1\psi \wedge \psi \to \neg\hat{\Gamma}'$. By applying the necessitation rule we get: $\vdash \Box_1\Box_1\psi \wedge \Box_1\psi \to \Box_1\neg\hat{\Gamma}'$ and by axiom 4 for $\Box_1$ we conclude $\vdash \Box_1\psi \to \Box_1\neg\hat{\Gamma}'$. Now as $\Box_1\psi \in \Gamma$, we have $\vdash \hat{\Gamma} \to \Box_1\neg\hat{\Gamma}'$ and as $\Gamma$ and $\Gamma'$ were taken arbitrarily we get $\vdash \bigvee_{\Gamma \in D} \hat{\Gamma} \to \bigwedge_{\Gamma' \notin D} \Box_1\neg\hat{\Gamma}'$. It is not difficult to prove that $\vdash \bigwedge_{\Gamma' \notin D} \Box_1\neg\hat{\Gamma} \leftrightarrow \Box_1 \bigvee_{\Gamma \in D} \hat{\Gamma}$, so we obtain the desired result $\vdash \delta \to \Box_1\delta$. Analogously we can prove $\vdash \delta \to \Box_2\delta$.

Now combining $\vdash \delta \to \Box_1\beta$ and $\vdash \delta \to \Box_1\delta$ yields $\vdash \delta \to \Box_1(\delta \wedge \beta)$ and analogously $\vdash \delta \to \Box_2(\delta \wedge \beta)$. So we have $\vdash \delta \to \Box_1(\delta \wedge \beta) \wedge \Box_2(\delta \wedge \beta)$. Now we apply the induction rule to obtain $\vdash \delta \to C_B\beta$. In particular we have $\vdash \hat{\Gamma} \to C_B\beta$. The last validity implies that $C_B\beta \in \Gamma$. So we have proved the left-to-right direction of the truth lemma for the case $\alpha = C_B\beta$.

For the other direction assume $C_B\beta \in \Gamma$. Let us show by induction on $k$ that if $\Gamma'$ is reachable from $\Gamma$ in $k$ steps then both $C_B\beta$ and $\beta$ are in $\Gamma'$.

Case for $k = 1$: Without loss of generality we can assume that $\Gamma R_1 \Gamma'$. By the axiom (*Equi*) we have $\vdash C_B\beta \rightarrow \Box_1\beta \wedge \Box_1 C_B\beta$. Now by construction both $\Box_1\beta, \Box_1 C_B\beta \in FL(\varphi)$. This implies that $\Box_1 C_B\beta \in \Gamma$ and $\Box_1\beta \in \Gamma$. By the definition of $R_1$ we get $\Gamma' \vdash \Box_1\beta \wedge \beta$ and $\Gamma' \vdash \Box_1 C_B\beta \wedge C_B\beta$. This implies that $\Gamma' \vdash \beta$ and $\Gamma' \vdash C_B\beta$ and as $\beta$ and $C_B\beta$ are in $FL(\varphi)$ we derive $\beta \in \Gamma'$ and $C_B\beta \in \Gamma'$.

Assume the induction hypothesis holds for $k \leq n$ and let us verify the case $k = n$. So we have $\Gamma R_x \Gamma_1 R_x \ldots R_x \Gamma_{n-1} R_x \Gamma'$, where $x \in \{1, 2\}$. By the induction hypothesis both $C_B\beta$ and $\beta$ are in $\Gamma_{n-1}$, so by the same argument as in the case of $k = 1$ we obtain $\beta \in \Gamma'$, hence $\Gamma \Vdash C_B\beta$. This finishes the truth lemma.

Now if we take $\Gamma_{\neg\varphi}$ to be a maximally consistent set containing $\neg\varphi$, by the truth lemma we it follows that $\mathcal{M}, \Gamma_{\neg\varphi} \nVdash \varphi$. This finishes the completeness proof.

We have seen that every non-theorem of $\mathbf{K4_2^C}$ is falsified on a finite, bi-transitive frame. The following theorem shows that every non-theorem of $\mathbf{K4_2^C}$ can be falsified on a frame $(W^t, R_1^t, R_2^t, V^t)$, where for each $k \in \{1, 2\}$ the pair $(W^t, R_k^t)$ is a transitive tree. Let us first recall the definition of tree.

**Definition 3.** A frame $(W, R)$ is called a *tree* if:

(1) it is rooted i.e., there is a unique point (the root) $r \in W$ such that for every $v \in W$ holds $v \neq r \Rightarrow rR^+v$,

(2) every element distinct from $r$ has a unique immediate predecessor; that is, for every $v \neq r$ there is a unique $v'$ such that $v'Rv$ and for every $v''$ we have that $v''Rv \Rightarrow v''Rv'$,

(3) $R$ is acyclic; that is, for every $v \in W$ we have $\neg vR^+v$.

If in addition $R$ is transitive i.e., $R = R^+$, then $(W, R)$ is called a transitive tree.

**Theorem 1.** *The modal logic* $\mathbf{K4_2^C}$ *has the tree model property.*

*Proof.* Suppose $\nvdash \varphi$. From Theorem 2 we know that $\varphi$ can be falsified in a finite, transitive, bi-relational Kripke model. Moreover, we can assume that this model is rooted. Let $\mathcal{M} = (W, R_1, R_2, V)$ be the model and $w$ be the root where $\varphi$ is falsified. Let us unravel the frame $(W, R_1, R_2)$ around $w$. As a result we get a frame $(W^t, R_1', R_2')$ where both $(W^t, R_1')$ and $(W^t, R_2')$ are trees. This is a standard technique in modal logic (Blackburn et al. 2006). The underlying set $W^t$ consists of all finite strings of the form $\langle w, w_1, \ldots, w_n\rangle$, where each $w_i \in W$ and $w(R_1 \cup R_2)w_1 \wedge w_i(R_1 \cup R_2)w_{i+1}$ for every $i \leq n-1$. The relation $R_k'$ ($k \in \{1, 2\}$) is defined in the following way: $\langle w, w_1, \ldots, w_n\rangle R_k' \langle w, w_1', \ldots, w_m'\rangle$ iff $m = n + 1$, $w_i = w_i'$ for every $i \leq n$ and $w_n R_k w_m$. To spell this out, one sequence is in the $R_k'$ relation with another if the second sequence takes the first sequence and adds as a tail an element which is an $R_k$-successor of the tail of the first sequence. The relation $R_k^t$ is defined as a transitive closures of $R_k'$ i.e., $R_k^t = (R_k')^+$ for each $k \in \{1, 2\}$. We define the model $\mathcal{M}^t = (W^t, R_1^t, R_2^t, V^t)$, where the valuation $V^t$ is defined by reflecting the valuation $V$, so $\langle w_1, \ldots, w_n\rangle \Vdash p$ iff $w_n \Vdash p$. It is easy to see that the function $f : W^t \rightarrow W$ which sends each element $\langle w_1, \ldots, w_n\rangle$ of $W^t$ to its tail $w_n$, is a bounded morphism from the model $\mathcal{M}^t = (W^t, R_1^t, R_2^t, V^t)$ to the model $\mathcal{M} = (W, R_1, R_2, V)$. At this point we can say that if $\varphi$ does not contain the common belief operator $C_B$ then $\mathcal{M}^t, w \nVdash \varphi$. This is because the bounded morphism preserves the satisfaction of formulas. But we can not yet say that the defined bounded morphism $f$ preserves formulas containing $C_B$. In fact it does. We can easily show that the function $f$ defined above is a bounded morphism between the extended models $\mathcal{M}^t = (W^t, R_1^t, R_2^t, (R_1^t \cup R_2^t)^+, V^t)$ and $\mathcal{M} = (W, R_1, R_2, (R_1 \cup R_2)^+, V)$.

*Note 1.* Observe that the relation $(R_1^t \cup R_2^t)^+$ does not contain cycles and in particular it is irreflexive. This is because if $\langle w, w_1, \dots, w_n\rangle (R_1^t \cup R_2^t)^+ \langle w, v_1, \dots, v_m\rangle$ then $m$ is strictly greater than $n$.

The main reason for introducing $\mathbf{K4_2^C}$ was to mimic the infinitary operator $C_B^\omega$ by finitary $C_B$. Though we cannot claim that on a logical level $C_B$ and $C_B^\omega$ are equivalent, we can establish a semantical equivalence, in particular on Kripke structures.

**Theorem 2.** *For any transitive bi-relational Kripke model $\mathcal{M} = (W, R_1, R_2, V)$ and point $w$: $\mathcal{M}, w \Vdash C_B\varphi$ iff $\mathcal{M}, w \Vdash C_B^\omega\varphi$.*

*Proof.* The proof follows easily from Definitions 1 and 2 inasmuch as both operators exactly depend on $(R_1 \cup R_2)$ – paths of finite length starting at $w$.

## 2.5 Common Belief as Equilibrium

We mentioned that common belief can also be understood as an equilibrium concept.[2] On Kripke structures the equilibrium concept coincides with common belief by infinite iteration, while in general the equilibrium concept has a much closer connection to the logic $\mathbf{K4_2^C}$. It can be formalized in the modal $\mu$-calculus in the following way:

$$C_\nu\varphi = \nu.p(\Box_1\varphi \wedge \Box_2\varphi \wedge \Box_1 p \wedge \Box_2 p).$$

The greatest fixpoint $\nu$ is defined as the fixpoint of a descending approximation sequence defined over the ordinals. Denote by $|\varphi|$ the truth set of $\varphi$ in the appropriate model $\mathcal{M}$ where evaluation occurs:

$$|C_\nu^0\varphi| = |\Box_1\varphi \wedge \Box_2\varphi|;$$
$$|C_\nu^{k+1}\varphi| = |\Box_1\varphi \wedge \Box_2\varphi \wedge \Box_1 C_\nu^k\varphi \wedge \Box_2 C_\nu^k\varphi|;$$

[2] For the remainder of this section and later on for Theorem 9 we assume some familiarity with the modal $\mu$-calculus. Lack of space hinders a fuller treatment, however for more details on the modal $\mu$-calculus we refer to Blackburn et al. (2006, Part 3, Chapter 4); see also the discussion in van Benthem and Sarenac (2004).

$$|C_\nu^\lambda \varphi| = |\bigcap_{k<\lambda} C_\nu^k \varphi|, \text{ for } \lambda \text{ a limit ordinal.}$$

We obtain $|C_\nu \varphi| = |C_\nu^\gamma \varphi|$, where $\gamma$ is a least ordinal for which the approximation procedure halts: i.e. $|C_\nu^\gamma \varphi| = |C_\nu^{\gamma+1} \varphi|$. Halting is guaranteed because the occurrence of the propositional variable $p$ in operator $F(p)$, where $F(p) = \Box_1 \varphi \wedge \Box_2 \varphi \wedge \Box_1 p \wedge \Box_2 p$, is positive. Hence by the Knaster-Tarski theorem the sequence will always reach a greatest fixpoint. Then the semantics of the operator $C_\nu$ is defined in the following way:

$$\mathcal{M}, w \Vdash C_\nu \varphi \text{ iff } w \in |C_\nu^\gamma \varphi|$$

In general this procedure may take more than $\omega$ steps, but in the case of Kripke structures the situation is simpler. The following property relates the different operators on Kripke models.

**Theorem 3.** *For every bi-relational Kripke model $\mathcal{M} = (W, R_1, R_2, V)$ and a point $w \in W$ the following condition holds: $\mathcal{M}, w \Vdash C_B^\omega \varphi$ iff $\mathcal{M}, w \Vdash C_\nu \varphi$.*

*Proof.* Observe that we can rewrite $C_B^\omega \varphi = \Box_1 \varphi \wedge \Box_2 \varphi \wedge \Box_1 \Box_1 \varphi \wedge \Box_1 \Box_2 \varphi \wedge \Box_2 \Box_1 \varphi \wedge \Box_2 \Box_2 \varphi \wedge \Box_1 \Box_1 \Box_1 \varphi \wedge \Box_1 \Box_1 \Box_2 \varphi \ldots$ in the following way: $\Box_1 \varphi \wedge \Box_2 \varphi \wedge \Box_1(\Box_1 \varphi \wedge \Box_2 \varphi) \wedge \Box_2(\Box_1 \varphi \wedge \Box_2 \varphi) \wedge \ldots$. Hence $|C_B^\omega \varphi| = |C_\nu^\omega \varphi|$. It is known that on Kripke structures the stabilization process does not need more than $\omega$ steps (van Benthem and Sarenac 2004) i.e. $|C_\nu \varphi| = |C_\nu^\omega \varphi|$. Hence $w \Vdash C_\nu \varphi$ iff $w \Vdash C_B^\omega \varphi$.

It follows that on transitive bi-relational Kripke structures the three operators $C_B$, $C_B^\omega$ and $C_\nu$ coincide.

## 2.6 A Note on the Semantics of Lismont and Mongin

In their paper (Lismont and Mongin 1994), Lismont and Mongin develop a neighborhood semantics for logics extended with a common belief operator. As a basis for the semantics they consider the class of augmented neighbourhood structures, i.e. the neighborhood function $N_i : W \to PP(W)$ for each agent $i \in \{1, 2\}$ has the following properties: for an arbitrary world $w \in W$, $N_i(w)$ contains the set $W$, it is closed under supersets and arbitrary intersections (the original work is presented for the finite set of agents we just simplify it here for the case of two agents). It is well known that there is a satisfaction preserving correspondence between augmented neighbourhood structures and Kripke structures and therefore one can reduce the completeness problem of a logic in neighborhood semantics to Kripke completeness (Chellas 1980; Hansen et al. 2009), although the main point is the definition of the semantics for the common belief operator in these terms. In the paper it is given by the following clause:

$$N_{C_\nu} = N_E \circ (N_{C_\nu} \cap B)$$

where $N_E$ stands for the semantics of the collective belief operator. In the case of two agents, $N_E(w) = N_1(w) \cap N_2(w)$ for every $w \in W$. The composition $\circ$ of neighborhood functions is defined in the following way $U \in (N \circ M)(w)$ iff $\{v \mid U \in M(v)\} \in N(w)$ for each $U \subseteq W$ and $w \in W$. Additionally $B$ is the neighborhood function defined by $U \in B(w)$ iff $w \in U$.

On the one hand we can see that $N_{C_v}$ is defined as a fixpoint, although it is not claimed to be the greatest fixpoint. On the other hand the definition of composition of neighborhood functions suggests the operator is treated as the infinite intersection of iterated modalities, exactly as in definition of the operator $C_B^{\omega}$. Therefore the definition of $N_{C_v}$ embraces both the fixpoint definition and the iteration of individual modalities at once and indeed on the class of augmented neighborhood structures these two definitions collapse to the two definitions of common belief operator on Kripke frames which we know to coincide. In general, however, the situation may be different. For example this is the case on the class of topological structures from van Benthem and Sarenac (2004).

In the next section we turn to the topological semantics for our common belief logic. It is natural to ask whether and how this is related to the neighbourhood semantics of Lismont and Mongin (1994). At present we do not have a precise answer to this question. One might look at topologies as special cases of neighborhood structures, where indeed neighborhoods are simply open neighborhoods of points in a topological sense. But this does not provide us with our derived set topological semantics, i.e. given a neighborhood model $(W, N, V)$ the truth set $\{w \mid \{v \in W \mid v \Vdash p\} \in N(w)\}$ of the modality $\Box p$ taken in the neighborhood semantics is not the same as the set of all colimits of the set $\{w \mid w \Vdash p\}$ in the topology obtained from the neighborhood function $N$. In fact the problem is that the class of neigbourhood structures that correspond to topological structures preserving the satisfaction of modal formulas has not yet been studied. Observe that here we deal with the derived set topological semantics, and we are supposing that neighborhood structures should preserve the satisfaction of formulas with respect to this d-semantics, and not with respect to the standard topological semantics.

## 3 Topological Semantics

The idea of a derived set topological semantics originates with the McKinsey-Tarski paper (McKinsey and Tarski 1944). This idea was taken further in Esakia (2001). The following works contain some important results in this direction: Bezhanishvil et al. (2005), Shehtman (1990), Lucero-Bryan (2011), and Gabelaia (2004). The derived set topological semantics for $\mathbf{K4_2^C}$ is provided by the class of all bitopological spaces. In the same way, as it is done in van Benthem and Sarenac (2004) for the common knowledge operator, we interpret the common belief operator on the intersection topology. On the other hand, different from $C_K$, for which the semantics is given using the interior of the intersection of the two topologies, we provide the semantics of $C_B\varphi$ as a set of all colimits of $|\varphi|$ in the intersection topology. As a main result we prove the soundness and completeness of the logic $\mathbf{K4}_{\mathbf{2}}^{\mathbf{C}}$ with respect to the class of all *$T_D$-intersection closed*, bi-topological spaces where each topology satisfies the $T_D$ separation axiom. We start with the basic definitions.

**Definition 4.** A pair $(X, \Omega)$ is called a topological space if $X$ is a set and $\Omega$ is a collection of subsets of $X$ with the following properties:

(1) $X, \varnothing \in \Omega$,
(2) $A, B \in \Omega$ implies $A \cap B \in \Omega$,
(3) $A_i \in \Omega$ implies $\bigcup A_i \in \Omega$.

Elements of $\Omega$ are called opens or open sets of the topological space.

**Definition 5.** A topological space $(X, \Omega)$ is called an *Alexandroff space* if an arbitrary intersection of opens is open, that is $A_i \in \Omega$ implies $\bigcap A_i \in \Omega$. $(X, \Omega)$ is called a *$T_D$-space* if every point $x \in X$ can be represented as an intersection of some open set $A$ and some closed set $B$.

We now define the colimit operator (or the set of all colimit points Engelking 1977) of a set in a topological space. This is needed to give the semantics of modal formulas in an arbitrary topological space.

**Definition 6.** Given a topological space $(X, \Omega)$ and a set $A \subseteq X$ we will say that $x \in X$ is a colimit point of $A$ if there exists an open neighborhood $U_x$ of $x$ such that $U_x - \{x\} \subseteq A$. The set of all colimit points of $A$ will be denoted by $\tau(A)$ and will be called the colimit set of $A$.

In words, a point $x$ belongs to the colimit points of a set $A$ iff some open set $B$ around $x$ is contained in $A \cup \{x\}$. The colimit set provides a semantics for the box modality, consequently the semantics for diamond is provided by the dual of the colimit set, which is called the *derived* set. The derived set of $A$ is denoted by $der(A)$. So we have $\tau(A) = X - der(X - A)$. Again a point $x$ belongs to the set of limit points of a set $A$ iff every open set $B$ around $x$ intersects with $A - \{x\}$. Below we list some examples and properties of the colimit and derivative operators.

*Example 1.* Let $R$ be a set of all reals and $A \subseteq R$ be as follows: $A = \{\frac{1}{m} \mid m \geq 1\}$. Then $der(A) = \{0\}$.

*Example 2.* Let $X$ be an arbitrary set and let $\Omega = \{U \mid U \subseteq X\}$, i.e. $\Omega$ is a discrete topology on $X$. Then for an arbitrary set $A \subseteq X$ we have the set of all colimit points $\tau(A)$ of a set $A$ is equal to $X$.

*Example 3.* Let $X$ be an arbitrary set and let $\Omega = \{\emptyset, X\}$, i.e. $\Omega$ is a trivial topology on $X$. Then for an arbitrary set $A \subseteq X$ the set of all colimit points $\tau(A)$ of $A$ is calculated as follows: If $X - A$ is a singleton or if $A = X$ then $\tau(A) = A$ otherwise $\tau(A) = \emptyset$.

**Fact 4 (Engelking 1977; Esakia 2004).** *For a given topological space $(X, \Omega)$ the following properties hold:*

*(1)* $Int(A) = \tau(A) \cap A \subseteq \tau\tau(A)$*, where Int denotes the interior operator,*
*(2)* $\tau(X) = X$ *and* $\tau(A \cap B) = \tau(A) \cap \tau(B)$*,*
*(3) If* $\Omega$ *is a* $T_d$*-space then* $\tau(A) \subseteq \tau\tau(A)$*,*
*(4) If* $\Omega_1 \subseteq \Omega_2$ *then* $\tau_1(A) \subseteq \tau_2(A)$ *where* $\tau_i$*,* $i \in \{1, 2\}$ *is a colimit operator of the corresponding topology* $\Omega_i$*.*

The following links $T_D$-spaces and irreflexive transitive relational structures. This result is a special case of a more general correspondence between weakly-transitive and irreflexive relational structures and all *Alexandroff* spaces (Esakia 2001).

**Fact 5 (Esakia 2004).** *There is a one-to-one correspondence between Alexandroff,* $T_D$*-spaces and transitive, irreflexive relational structures.*

Let us briefly describe the correspondence. We first introduce the downset operator. Let $(X, R)$ be a Kripke frame. The downset operator $R^{-1}$ is defined in the following way: for any $A \subseteq X$ we set $R^{-1}(A) := \{x|(\exists y)(y \in A \wedge xRy)\}$. Now if we are given an irreflexive, transitive order $(X, R)$ it is possible to prove that the downset operator $R^{-1}$ satisfies all the properties of the topological derivative operator for $T_D$-spaces. Hence we get a $T_D$-space $(X, \Omega_R)$, where $\Omega_R$ is the topology obtained from the derivative operator $R^{-1}$. Conversely with every *Alexandroff* $T_D$-space $(X, \Omega)$, one can associate an irreflexive and transitive relational structure $(X, R_\Omega)$, where $xR_\Omega y$ iff $x \in der(\{y\})$. Moreover we have that $(X, \Omega_{R_\Omega})$ is homeomorphic to $(X, \Omega)$ and $(X, R_{\Omega_R})$ is order isomorphic to $(X, R)$.

**Fact 6 (Esakia 2004).** *The set* $A$ *is open in* $(X, \Omega_R)$ *iff* $x \in A$ *implies that the implication* $(xRy \Rightarrow y \in A)$ *holds for every* $y \in X$*.*

This correspondence can be directly generalized to Kripke frames with more than one transitive and irreflexive relation. Of course then we will have one Alexandroff $T_D$-space for each irreflexive and transitive order. Below we prove the proposition which builds a bridge between Kripke and topological semantics for $\mathbf{K4_2^C}$.

**Proposition 3.** *If* $R_1$ *and* $R_2$ *are two irreflexive and transitive orders on* $X$ *and* $(R_1 \cup R_2)^+$ *is also irreflexive and transitive, then* $\Omega_{(R_1 \cup R_2)^+} \cong \Omega_{R_1} \cap \Omega_{R_2}$*.*

Before starting the proof, observe that $(R_1 \cup R_2)^+$ may not be irreflexive even if both $R_1$ and $R_2$ are. For example: let $X = \{x, y\}$ and $R_1 = \{(x, y)\}$ and $R_2 = \{(y, x)\}$ then $(R_1 \cup R_2)^+ = \{(x, y), (y, x), (x, x), (y, y)\}$. On the topological side this example shows that $T_D$-spaces do not form a lattice. That is why in Proposition 3 we require $(R_1 \cup R_2)^+$ to be irreflexive and transitive.

*Proof.* Assume that $A \in \Omega_{(R_1 \cup R_2)^+}$. By Fact 6 this means that if $x \in A$ then for every $y$ such that $x(R_1 \cup R_2)^+ y$ it holds that $y \in A$. Since $R_i \subseteq (R_1 \cup R_2)^+$ for each $i \in \{1, 2\}$, it holds that $xR_1 y \Rightarrow y \in A$ and $xR_2 y \Rightarrow y \in A$ for every $y \in X$. Hence $A \in \Omega_1 \cap \Omega_2$ according to Fact 6.

Conversely assume $A \in \Omega_1 \cap \Omega_2$. This means that $x \in A \Rightarrow (x(R_1 \cup R_2)y \Rightarrow y \in A)$. Now take an arbitrary $y$ such that $x(R_1 \cup R_2)^+ y$. By definition this means that there is a $(R_1 \cup R_2)$-path $\langle x_1, x_2, \ldots x_n\rangle$ starting at $x$ going to $y$. But this means that each member of this path is in $A$ because $A$ is open in the intersection of the two topologies. Hence $y \in A$ and hence $A \in \Omega_{(R_1 \cup R_2)^+}$

Next we give a definition of the satisfaction relation of modal formulas in the derived set topological semantics. Observe that this definition is given in a standard modal language i.e., without the common belief operator. Recall that a topological model is a tuple $\mathcal{M} = (W, \Omega, V)$ where $V : Prop \to P(W)$ is a valuation function.

**Definition 7.** The satisfaction of a modal formula in a topological model $\mathcal{M} = (W, \Omega, V)$ at a point $w \in W$ is defined in the following way:

- $\mathcal{M}, w \Vdash p$ iff $w \in V(p)$,
- Boolean cases are standard,
- $\mathcal{M}, w \Vdash \Box\varphi$ iff $w \in \tau(V(\varphi))$, where $\tau$ is a colimit operator of $\Omega$.

**Fact 7 (Esakia 2004).** *The correspondence mentioned in Fact 5 preserves the truth of modal formulas, i.e.* $(W, R, V), x \Vdash \alpha$ *iff* $(W, \Omega_R, V), x \Vdash \alpha$.

Note that in Fact 7, the symbol $\Vdash$ on the left hand side denotes the satisfaction relation on Kripke models, while on the right hand side it denotes the satisfaction relation on topological frames in the derived set semantics. Now we extend the satisfaction relation to the language with the common belief operator.

**Definition 8.** The satisfaction of a modal formula on a bi-topological model $\mathcal{M} = (W, \Omega_1, \Omega_2, V)$ at a point $w \in W$ is defined in the following way:

$\mathcal{M}, w \Vdash p$ iff $w \in V(p)$,
$\mathcal{M}, w \Vdash \alpha \wedge \beta$ iff $\mathcal{M}, w \Vdash \alpha$ and $\mathcal{M}, w \Vdash \beta$,
$\mathcal{M}, w \Vdash \neg\alpha$ iff $\mathcal{M}, w \nVdash \alpha$,
$\mathcal{M}, w \Vdash \Box_i\varphi$ iff $w \in \tau_i(V(\varphi))$, where $\tau_i$ is a colimit operator of $\Omega_i$, $i \in \{1, 2\}$,
$\mathcal{M}, w \Vdash C_B\varphi$ iff $w \in \tau_{1\wedge 2}(V(\varphi))$, where $\tau_{1\wedge 2}$ is a colimit operator in $\Omega_1 \cap \Omega_2$.

As an immediate corollary of Proposition 3 and a many-modal version of Fact 7, we get the following proposition.

**Proposition 4.** *If* $R_1$ *and* $R_2$ *are two irreflexive and transitive orders and* $(R_1 \cup R_2)^+$ *is also topological then for every formula* $\alpha$ *in* $\mathbf{K4_2^C}$ *the following holds:*

$$(W, R_1, R_2, V), x \Vdash \alpha \text{ iff } (W, \Omega_{R_1}, \Omega_{R_2}, V), x \Vdash \alpha.$$

Now it is clear that we can reduce the topological completeness problem to Kripke completeness if for every non-theorem $\mathbf{K4_2^C} \nvdash \varphi$ we can find a bi-relational topological counter-model $(W, R_1, R_2, V)$ with $(R_1 \cup R_2)^+$ being also a topological relation.

**Definition 9.** The triple $(X, \Omega_1, \Omega_2)$ is a $T_D$-intersection closed bi-topological space if each of the topologies $\Omega_1$, $\Omega_2$ and $\Omega_1 \cap \Omega_2$, satisfies the $T_D$-separation axiom.

**Theorem 8.** $\mathbf{K4_2^C}$ *is sound and complete with respect to the class of all $T_D$-intersection closed, bi-topological, Alexandroff spaces.*

*Proof.* (Soundness) Take an arbitrary $T_D$-intersection closed, bi-topological model $\mathcal{M} = (X, \Omega_1, \Omega_2, V)$. From (2) and (3) of Fact 4 it follows that $K4$-axioms are valid for each box. Let us show that at each point $x \in X$, the equilibrium axiom is satisfied. Assume that $\mathcal{M}, x \Vdash C_B p$. Hence by Definition 8 we have $x \in \tau_{1\wedge 2}|p|$. By (4) of Fact 4 we get $x \in \tau_1|p|$ and $x \in \tau_2|p|$. By (3) we have $\tau_{1\wedge 2}|p| \subseteq \tau_{1\wedge 2}\tau_{1\wedge 2}|p| \subseteq \tau_1\tau_{1\wedge 2}|p|$. Analogously $\tau_{1\wedge 2}|p| \subseteq \tau_2\tau_{1\wedge 2}|p|$. Hence we have $x \Vdash \Box_1 p \wedge \Box_2 p \wedge \Box_1 C_B p \wedge \Box_2 C_B p$.

For the other direction assume that $x \in \tau_1\tau_{1\wedge 2}|p| \cap \tau_1|p| \cap \tau_2\tau_{1\wedge 2}|p| \cap \tau_2|p|$. By (2) of Fact 4 we get $x \in \tau_1(\tau_{1\wedge 2}|p| \cap |p|) \cap \tau_2(\tau_{1\wedge 2}|p| \cap |p|)$. By (1) of Fact 4 we conclude $x \in \tau_1(Int_{1\wedge 2}|p|) \cap \tau_2(Int_{1\wedge 2}|p|)$, where $Int_{1\wedge 2}$ denotes the interior operator in the intersection topology. By the definition of colimit there exists $U_x^1 \in \Omega_1$ such that $x \in U_x^1$ and $U_x^1 - \{x\} \subseteq Int_{1\wedge 2}|p|$ and there exists $U_x^2 \in \Omega_2$ such that $x \in U_x^2$ and $U_x^2 - \{x\} \subseteq Int_{1\wedge 2}|p|$. Hence $(U_x^1 \cup U_x^2) - \{x\} \subseteq Int_{1\wedge 2}|p|$. Let us show that $Int_{1\wedge 2}|p| \cup \{x\}$ is open in $\Omega_1 \cap \Omega_2$. Since $U_x^1 \in \Omega_1$ and $Int_{1\wedge 2}|p| \in \Omega_1$ we have $U_x^1 \cup Int_{1\wedge 2}|p| = Int_{1\wedge 2}|p| \cup \{x\} \in \Omega_1$. Analogously we show that $Int_{1\wedge 2}|p| \cup \{x\} \in \Omega_2$. Hence $x \in \tau_{1\wedge 2}|p|$.

Let us show that the induction rule is valid in the class of all $T_D$-intersection closed bi-topological spaces. The proof goes by contraposition. Assume not $\vdash p \to C_B q$. This means that for some $T_D$-intersection closed, bi-topological model $\mathcal{M} = (X, \Omega_1, \Omega_2, V)$ and a point $x \in X$ it holds that: $x \Vdash p$ while $x \nVdash C_B q$. We want to show that not $\vdash p \to \Box_1(p \wedge q) \wedge \Box_2(p \wedge q)$. It suffices to find a $T_D$-intersection closed bi-topological model which falsifies the formula. For such a model one could take $\mathcal{M}' = (X, \Omega_1 \cap \Omega_2, \Omega_1 \cap \Omega_2, V)$. Indeed as $(X, \Omega_1, \Omega_2, V)$ is $T_D$-intersection closed, the topology $\Omega_1 \cap \Omega_2$ satisfies the $T_D$-separation axiom. Besides since in $\mathcal{M}'$ both topologies are the same, their intersection is also $\Omega_1 \cap \Omega_2$ and hence again is a $T_D$-space. Now it is immediate that $\mathcal{M}', x \nVdash p \to \Box_1(p \wedge q) \wedge \Box_2(p \wedge q)$. This is because by construction of $\mathcal{M}'$ we have $\mathcal{M}', x \nVdash \Box_i q$ iff $\mathcal{M}, x \nVdash C_B q$ for every $x \in X$ and $i \in \{1, 2\}$.

(Completeness) Assume $\mathbf{K4_2^C} \nvdash \varphi$. According to Theorem 1 there exist a tree model $M^t = (W^t, R_1^t, R_2^t, V)$ which falsifies $\varphi$. We know that $(R_1 \cup R_2)^+$ is an irreflexive and transitive order (see Note 1). By applying Proposition 4 it follows that the formula $\varphi$ is falsified in the corresponding bi-topological model $(W^t, \Omega_{R_1^t}, \Omega_{R_2^t}, V)$, which is $T_D$-intersection closed because of Fact 5, Proposition 3 and Note 1.

We can now show how the semantical definition of common belief $C_B\varphi$ as a colimit of the intersection topology meshes with the general equilibrium concept: on topological models the two operators $C_B$ and $C_\nu$ coincide.

**Theorem 9.** *For every bi-topological model $\mathcal{M} = (X, \Omega_1, \Omega_2, V)$ and an arbitrary formula $\varphi$ the following equality holds: $\nu.p(\tau_1(|\varphi|) \cap \tau_2(|\varphi|) \cap \tau_1(p) \cap \tau_2(p)) = \tau_{1\wedge 2}(|\varphi|)$.*

*Proof.* That $\tau_{1\wedge 2}(|\varphi|)$ is a fixpoint of the operator $F(p) = \tau_1(|\varphi|) \cap \tau_2(|\varphi|) \cap \tau_1(p) \cap \tau_2(p)$ follows from the soundness proof of the equilibrium axiom, see Theorem 8.

Now let us show that $\tau_{1\wedge 2}(|\varphi|)$ is the greatest fixpoint of $F(p)$. Take an arbitrary fixpoint $B$ of the operator $F(p)$. That $B$ is a fixpoint immediately implies that $B \subseteq \tau_1(|\varphi|) \cap \tau_2(|\varphi|) \cap \tau_1(B) \cap \tau_2(B)$. By (1) of Fact 4 we have $B \subseteq Int_i(B) = \tau_i(B) \cap B$ for each $i \in \{1,2\}$. Hence $B = Int_{1\wedge 2}(B)$ where $Int_{1\wedge 2}$ is the interior operator in the intersection topology of the two topologies. Now let us show that for every $x \in B$ the set $\{x\} \cup (B \cap |\varphi|)$ is open in the intersection of the two topologies. Take an arbitrary point $y \in \{x\} \cup (B \cap |\varphi|)$. Since $y \in B \subseteq \tau_1(|\varphi|)$ we know that there exists an open neighborhood $U^1_y \in \Omega_1$ of $y$ such that $U^1_y - \{y\} \subseteq |\varphi|$. This means that $B \cap U^1_y \in \Omega_1$ and $B \cap U^1_y \subseteq \{x\} \cup (B \cap |\varphi|)$. This means that for every point $y \in \{x\} \cup (B \cap |\varphi|)$ there is an open neighborhood $B \cap U^1_y \in \Omega_1$ of $y$ such that $B \cap U^1_y \subseteq \{x\} \cup (B \cap |\varphi|)$ hence $\{x\} \cup (B \cap |\varphi|) \in \Omega_1$. In exactly the same way we show that $\{x\} \cup (B \cap |\varphi|) \in \Omega_2$. Hence $\{x\} \cup (B \cap |\varphi|) \in \Omega_1 \cap \Omega_2$. This means that $x \in \tau_{1\wedge 2}(|\varphi|)$ since there exists an open neighborhood $U_{1\wedge 2} = \{x\} \cup (B \cap |\varphi|) \in \Omega_1 \cap \Omega_2$ with $U_{1\wedge 2} - \{x\} \in |\varphi|$.

## 4 From Belief to Knowledge

Let us now look briefly at the connection between the logics of common knowledge $\mathbf{S4^C_2}$ and common belief $\mathbf{K4^C_2}$. This connection generalizes the existing splitting translation between **S4**-logics and **K4**-logics.[3] As a result we obtain a validity preserving translation from $\mathbf{S4^C_2}$ formulas to $\mathbf{K4^C_2}$ formulas in which common knowledge is expressed in terms of common belief.

**Definition 10.** The normal modal logic $\mathbf{S4^C_2}$ is defined in a modal language with infinite set of propositional letters $p, q, r \ldots$ and connectives $\vee, \wedge, \neg, \Box_1, \Box_2, C_K$, where the formulas are constructed in a standard way.

- The axioms are all classical tautologies, each box satisfies all **S4** axioms and in addition we have the equilibrium axiom for the common knowledge operator:

$$(equi) : C_K p \leftrightarrow p \wedge \Box_1 C_K p \wedge \Box_2 C_K p$$

- The rules of inference are: Modus-ponens, Substitution, Necessitation for $\Box_1$ and $\Box_2$ and the induction rule:

$$(ind) : \frac{\vdash \varphi \to \Box_1(\varphi \wedge \psi) \wedge \Box_2(\varphi \wedge \psi)}{\vdash \varphi \to C_K \psi}$$

  for arbitrary formulas $\varphi$ and $\psi$ of the language.

The Kripke semantics for the modal logic $\mathbf{S4^C_2}$ is provided by reflexive and transitive, bi-relational Kripke frames. To interpret the common knowledge operator $C_K$, the reflexive, transitive closure of a union relation is used.

[3] For a discussion of the splitting translation and its application in non-monotonic modal logics, see the authors' (Pearce and Uridia 2011a).

**Definition 11.** The reflexive, transitive closure $R^\star$ of a relation $R \subseteq W \times W$ is defined in the following way: $R^\star = R^+ \cup \{(w,w) | w \in W\}$.

The satisfaction of formulas is defined as follows.

**Definition 12.** For a given bi-relational Kripke model $\mathcal{M} = (W, R_1, R_2, V)$ the satisfaction of a formula at a point $w \in W$ is defined inductively as follows:

$w \Vdash p$ iff $w \in V(p)$,
$w \Vdash \alpha \wedge \beta$ iff $w \Vdash \alpha$ and $w \Vdash \beta$,
$w \Vdash \neg\alpha$ iff $w \nVdash \alpha$,
$w \Vdash \Box_i\varphi$ iff $(\forall v)(wR_iv \Rightarrow v \Vdash \varphi)$,
$w \Vdash C_K\varphi$ iff $(\forall v)(w(R_1 \cup R_2)^\star v \Rightarrow v \Vdash \varphi)$.

**Fact 10 (Fagin et al. 1995).** *The modal logic* $\mathbf{S4_2^C}$ *is sound and complete with respect to the class of all finite, reflexive, bi-transitive Kripke frames.*

**Definition 13.** Consider the following function from the set of formulas in $\mathbf{S4_2^C}$ to the set of formulas in $\mathbf{K4_2^C}$.

$Sp(p) = p$ for every propositional letter $p$,
$Sp(\neg\alpha \vee \beta) = \neg Sp(\alpha) \vee Sp(\beta)$,
$Sp(\Box_i\alpha) = \Box_i Sp(\alpha) \wedge Sp(\alpha)$,
$Sp(C_K\alpha) = C_B Sp(\alpha) \wedge Sp(\alpha)$.

**Theorem 11.** $\vdash_{\mathbf{S4_2^C}} \varphi$ *iff* $\vdash_{\mathbf{K4_2^C}} Sp(\varphi)$.

*Proof.* We prove the theorem by a semantical argument using the Kripke completeness results, see Proposition 2 and Fact 10. Let us first show by induction on the length of a formula that for every bi-relational Kripke model $\mathcal{M} = (W, R_1, R_2, V)$ and every $w \in W$ the following holds:

$$(a) \quad \mathcal{M}^\star = (W, R_1^\star, R_2^\star, V), w \Vdash \varphi \;\text{ iff }\; \mathcal{M}^+ = (W, R_1^+, R_2^+, V), w \Vdash Sp(\varphi).$$

The only nonstandard case is when $\varphi = C_K\psi$. Assume $\mathcal{M}^\star, w \Vdash C_K\psi$. By the definition of $(R_1 \cup R_2)^\star$ this means that $\mathcal{M}^\star, w \Vdash \psi$ and for every $w'$ such that $w(R_1 \cup R_2)^\star w'$, we have $\mathcal{M}^\star, w' \Vdash \psi$. Now by the induction hypothesis we have that $\mathcal{M}^+, w \Vdash \psi$ and $\mathcal{M}^+, w' \Vdash \psi$. Since $w'$ was an arbitrary $(R_1 \cup R_2)^\star$ successor of $w$ we have $\mathcal{M}^+, w \Vdash C_B\psi$. This is because $(R_1 \cup R_2)^\star \supseteq (R_1 \cup R_2)^+$. Hence we obtain $\mathcal{M}^+, w \Vdash C_B\psi \wedge \psi$. The converse direction follows by the same argument.

Now assume $\vdash_{\mathbf{S4_2^C}} \varphi$. By Fact 10 this means that $\varphi$ is valid in every reflexive and transitive, bi-relational model. Take an arbitrary transitive, bi-relational model $\mathcal{M}$. Then by assumption we have $\mathcal{M}^\star \Vdash \varphi$. Hence by $(a)$ we have that $\mathcal{M} \Vdash Sp(\varphi)$. As $\mathcal{M}$ was an arbitrary transitive, bi-relational model, from Proposition 2 we infer that $\vdash_{\mathbf{K4_2^C}} Sp(\varphi)$. Conversely, suppose $\vdash_{\mathbf{K4_2^C}} Sp(\varphi)$. Then by Proposition 2, $Sp(\varphi)$ is valid in the class of all transitive, bi-relational models. Take an arbitrary reflexive and transitive, bi-relational model $\mathcal{N}$. Then $\mathcal{N} \Vdash Sp(\varphi)$ because $\mathcal{N} = \mathcal{N}^+$. So by $(a)$ we have that $\mathcal{N}^\star \Vdash \varphi$. Now as $\mathcal{N}$ was reflexive and transitive, $\mathcal{N}^\star = \mathcal{N}$, hence $\mathcal{N} \Vdash \varphi$. And since $\mathcal{N}$ was an arbitrary reflexive and transitive, bi-relational model, by Fact 10 we have $\vdash_{\mathbf{S4_2^C}} \varphi$.

## 5 Conclusions

Our main aim in this chapter has been to extend the work of van Benthem and Sarenac (2004) on the topological semantics for common knowledge by interpreting a common belief operator on the intersection of two topologies in a bi-topological model. In particular we considered a logic $\mathbf{K4_2^C}$ of common belief for normal agents, first under a Kripke, relational semantics, showing it to have the finite model property and the tree model property. We then showed that $\mathbf{K4_2^C}$ is the modal logic of all $T_D$-intersection closed, bi-topological spaces with a derived set interpretation of modalities and we saw how the common knowledge logic $\mathbf{S4_2^C}$ can be embedded in $\mathbf{K4_2^C}$ via the splitting translation that maps $C_K p$ into $p \wedge C_B p$.

A worthwhile exercise for the future would be to undertake a more detailed comparison of our topological approach with the neighborhood systems of Lismont and Mongin (1994) that we became aware of after finishing the first version of this chapter. Another direction for the future would be to look for concrete topological structures which would fully capture the behavior of the logic $\mathbf{K4_2^C}$ or some of its extensions.

**Acknowledgements** An earlier version of this chapter was presented at the conference on Agreement Technologies, AT2012, 15–16 October 2012, Dubrovnik, Croatia. The authors are grateful to anonymous reviewers whose comments helped to improve readability. This research has been partially supported by the Spanish Ministry of Science and Innovation through the AT project CSD2007-0022 and MCICINN project TiN2009-14562-CO5, by the Shota Rustaveli National Science Foundation project 52/05 (PG/72/4-102/13) and by SINTELNET, the European Network for Social Intelligence.

## References

Aumann, Robert. 1976. Agreeing to disagree. *Annals of Statistics* 4(6): 1236–1239.

Baltag, Alexandru, and Sonja Smets. 2009. Group belief dynamics under iterated revision: Fixed points and cycles of joint upgrades. In *TARK*, New York, 41–50.

Baltag, Alexandru, Lawrence S. Moss, and Slawomir Solecki. 1998. The logic of public announcements, common knowledge, and private suspicions. In *Proceedings of the TARK'98*, Evanston. Morgan Kaufmann.

Barwise, Jon. 1988. Three views of common knowledge. In *Proceedings of the second conference on theoretical aspects of reasoning about knowledge*, 365–378. San Francisco: Morgan Kaufmann.

Bezhanishvili, Nick, and Wiebe van der Hoek. (2014). Structures for epistemic logic (survey). In *Logical and informational dynamics*, a volume in honour of Johan van Benthem, trends in logic, ed. A. Baltag and S. Smets. Springer.

Bezhanishvil, Guram, Leo Esakia, and David Gabelaia. 2005. Some results on modal axiomatization and definability for topological spaces. *Studia Logica* 81(3): 325–355.

Blackburn, Patrick, Johan van Benthem, and Frank Wolter. 2006. *Handbook of modal logic*. Amsterdam: Elsevier Science & Technology.

Chellas, Brian F. 1980. *Modal logic – An introduction*, 295 p. Cambridge/New York: Cambridge University Press.

Engelking, Ryszard. 1977. *General topology*. Warsaw: Taylor & Francis.

Esakia, Leo. 2001. Weak transitivity – Restitution. *Logical Studies* 8: 244–255.

Esakia, Leo. 2004. Intuitionistic logic and modality via topology. *Annals of Pure Applied Logic* 127(1–3): 155–170.

Fagin, Ronald, Joseph Y. Halpern, Yoram Moses, and Moshe Y. Vardi. 1995. *Reasoning about knowledge*. Cambridge: MIT Press.

Fischer, M.J., and R.E. Ladner. 1979. Propositional dynamic logic of regular programs. *Journal of Computer Sciences* 18: 194–211.

Gabelaia, David. 2004. Topological, algebraic and spati-temporal semantics for multi-dimentional modal logics. PhD thesis, King's College, London.

Hansen, Helle Hvid, Clemens Kupke, and Eric Pacuit. (2009). Neighbourhood structures: Bisimilarity and basic model theory. *Logical Methods in Computer Science* 5(2).

Herzig, Andreas, Tiago De Lima, and Emiliano Lorini. 2009. On the dynamics of institutional agreements. *Synthese* 171(2): 321–355.

Lewis, David. 1969. *Convention: A philosophical study*. Cambridge: Harvard University Press.

Lismont, Luc, and Philippe Mongin. 1994. On the logic of common belief and common knowledge. *Theory and Decision* 37: 75–106.

Lucero-Bryan, Joel. 2011. The *d*-logic of rational numbers: A new proof. *Studia Logica – An International Journal for Symbolic Logic – SLOGICA* 97(2): 265–295.

McKinsey, Jon, and Alfred Tarski. 1944. The algebra of topology. *Annals of Mathematics* 45: 141–191.

Pearce, David, and Levan Uridia. 2010. Minimal knowledge and belief via minimal topology. In *Logics in artificial intelligence, Proceedings of JELIA 2010*, LNAI 6341, ed. T. Janhunen, I. Niemela, 273–285. Berlin: Springer.

Pearce, David, and Levan Uridia. 2011a. The Gödel and the splitting translations. In *Nonmonotonic reasoning*, Studies in Logic, vol. 31, ed. G. Brewka, V. Marek, and M. Truszczynski, 335–360. London: College Publications.

Pearce, David, and Levan Uridia. 2011b. An approach to minimal belief via objective belief. In *Proceedings of the 22nd international joint conference on artificial intelligence, IJCAI 11*, Barcelona, ed. T. Walsh, 1045–1050.

Shehtman, Valentin. 1990. *Derived sets in Euclidean spaces and modal logic*. Amsterdam: University of Amsterdam. X-1990-05.

Smullyan, Raymond. 1986. Logicians who reason about themselves. In *Proceedings of the conference on theoretical aspects of reasoning about knowledge*, 341–352. San Francisco: Morgan Kaufmann.

Stalnaker, Robert. Common ground. 2001. *Linguistics and Philosophy* 25(5–6): 701–721.

Steinsvold, Christopher. 2008. A grim semantics for logics of belief. *Journal of Philosophical Logic* 37: 45–56.

Steinsvold, Christopher. 2009. Topological models of belief logics. PhD theis, VDM Verlag, Aug 27.

van Benthem, Johan. 2007. Rational dynamics and epistemic logic in games. *International Game Theory Review* 9: 13–45

van Benthem, Johan, and Darko Sarenac. 2004. The geometry of knowledge. In *Aspects of universal logic*, Travaux de logique, vol. 17, eds. J.-Y. Beziau, A. Costa Leite, and A. Facchini, 1–31. Neuchâtel: Centre de recherches sémiologiques/Université de Neuchâtel.

# Social Emotions from the Perspective of the Computational Belief-Desire Theory of Emotion

**Rainer Reisenzein**

**Abstract** At the center of the social emotions are reactions to the positive and negative fate of others and to the perceived fulfillment and violation of social and moral norms. Using pity and guilt as representatives of these two groups of social emotions, I investigate their generation, nature and function from the perspective of CBDTE, a (sketch of a) Computational model of the Belief-Desire Theory emotion. The central assumption of CBDTE is that a core subset of human emotions are the products of hardwired mechanisms whose primary function is to subserve the monitoring and updating of the belief-desire system. The emotion mechanisms work like sensory transducers; however, instead of sensing the world, they monitor the belief-desire system and signal important changes in this system, in particular the fulfillment and frustration of desires and the confirmation and disconfirmation of beliefs. Social emotions are accommodated into CBDTE by assuming that the proximate beliefs and desires that cause them are derived from special kinds of desire. Specifically, pity is a form of displeasure that is experienced if an altruistic desire is frustrated by the negative fate of another person; whereas guilt is a form of displeasure that is experienced if a nonegoistic desire to comply with a norm is frustrated by an own action. The intra-system function of these emotions is to signal the frustration of altruistic desires (pity) and of nonegoistic desires to comply with a norm (guilt) to other cognitive subsystems, to globally prepare and motivate the agent to deal with them. The communication of social emotions serves to reveal the person's social (nonegoistic) desires to others: Her altruistic concern for others, and her nonegoistic caring for the observance of social norms.

Parts of this article are based on a German book chapter (Reisenzein 2010). I am grateful to Cristiano Castelfranchi for his helpful comments to a previous version of the manuscript.

R. Reisenzein (✉)
Institute of Psychology, University of Greifswald, Franz-Mehring-Straße 47, 17487 Greifswald, Germany
e-mail: rainer.reisenzein@uni-greifswald.de

A. Herzig, E. Lorini (eds.), *The Cognitive Foundations of Group Attitudes and Social Interaction*, Studies in the Philosophy of Sociality 5,
DOI 10.1007/978-3-319-21732-1_8

**Keywords** Belief-desire theory of emotion • Computational modelling of emotions • Social emotions • Pity • Guilt

"Social emotions" can be roughly defined as emotions whose elicitors and objects essentially involve social agents (other persons, groups, institutions). Some social emotions, such as love and hate, attraction and repulsion, trust and distrust seem to have social agents themselves as objects; whereas others, such as anger about another's norm-violation or envy of another's good fortune, are directed at propositions or states of affairs that involve social agents. In this article, I am concerned with these "propositional" social emotions. At their core are two emotion families: the fortune-of-others emotions (Ortony et al. 1988), i.e. emotional reactions to the positive or negative fate of other people, such as joy for another, envy, pity and Schadenfreude (gloating); and the norm-based emotions, i.e. emotional reactions to the perceived violation and fulfillment of social and moral norms, such as guilt, shame, indignation and moral elevation. In this article, I investigate the generation, nature and function of these two kinds of social emotions from the perspective of CBDTE (Reisenzein 2009a, b; also see Reisenzein 2001, 2012a, b; Reisenzein and Junge 2012), a (sketch of a) computational (C) model of the belief-desire theory of emotion (BDTE). In Section 1, I summarize CBDTE. In Section 2, I discuss how CBDTE explains the social emotions, using the examples of pity and guilt.

## 1 The Computational Belief-Desire Theory of Emotion

The starting point of the computational model of emotion sketched in Reisenzein (2009a, b; see also, Reisenzein 2001) is the *cognitive-motivational*, or *belief-desire theory of emotion* (BDTE). BDTE, in turn, is a member of the family of cognitive emotion theories that have dominated discussions of emotions during the past 30 years in both psychology and philosophy (for reviews, see e.g., Ellsworth and Scherer 2003; Goldie 2007). As explained below, BDTE differs from the standard version of cognitive emotion theory (the cognitive-evaluative theory of emotion) in a number of foundational assumptions that allow BDTE to escape several criticisms of the standard view; or at least so its proponents argue. Although BDTE has been primarily promoted by philosophers (see especially Davis 1981; Green 1992; Marks 1982; Searle 1983), it also has adherents in psychology (e.g., Castelfranchi and Miceli 2009; Oatley 2009; Reisenzein 2001, 2009a; Roseman 1979). Recent formal reconstructions of cognitive emotion theories (e.g., Adam et al. 2009; Steunebrink et al. 2012) have also adopted the belief-desire framework (see Reisenzein et al. 2013).

The most important difference between BDTE and the standard version of cognitive emotion theory concerns what a pioneer BDTE theorist, the Austrian philosopher-psychologist Alexius Meinong (1894), called the "psychological preconditions" of emotions: the mental states required for having an emotion. According to the standard version of cognitive emotion theory, the *cognitive-evaluative* theory of emotion—known as *appraisal theory* in psychology (e.g., Arnold 1960; Frijda 1986; Lazarus 1991; Scherer 2001) and as the *judgment theory* of emotions in philosophy (e.g., Solomon 1976; Nussbaum 2001)—emotions presuppose certain factual and evaluative cognitions about their eliciting events, which in their paradigmatic form are factual and evaluative beliefs. In contrast, BDTE is a *cognitive-motivational* theory of emotion: It assumes that emotions depend not only on beliefs (i.e., cognitive or informational states) but also on desires (i.e., motivational states) (for an elaboration of the distinction between beliefs and desires see e.g., Green 1992; Smith 1994).

To illustrate the difference between the two theories, assume that Maria feels happy that *Mr. Schroiber was elected chancellor.* According to the cognitive-evaluative theory of emotion, Maria experiences happiness about this state of affairs *p* only if, and under "normal working conditions" always if (see Reisenzein 2012a), she comes to (firmly) believe that *p* obtains, and evaluates *p* as good for herself (i.e., believes that *p* is good for her). In contrast, according to BDTE, Maria feels happy about *p* if she comes to believe *p*, and if she desires *p*. Although many proponents of the theory (including most psychological appraisal theorists) acknowledge that desires are also important for emotions, inasmuch as appraisals of events express their relevance for the person's motives, desires, or goals (e.g., Lazarus 1991; Scherer 2001; Ortony et al. 1988), the link between desires and emotions is held to be mediated by appraisals (Reisenzein 2006a). In contrast, according to BDTE, emotions are based *directly* on desires and (typically factual) beliefs (Green 1992; Reisenzein 2009b; see, also Castelfranchi and Miceli 2009). Although this difference between the two theories may at first sight appear to be small, it has a profound implication: It implies that the evaluative cognitions that are at the center of the cognitive-evaluative theory are in fact neither necessary nor, together with factual beliefs, sufficient for emotions. All that is needed for feeling happy about *p* is desiring *p* and believing that *p* obtains. It is not necessary to, in addition, believe that *p* is good for oneself, or fulfills a desire.

It should be noted that in contrast to other belief-desire theorists (e.g., Castelfranchi and Miceli 2009; Marks 1982; Green 1992), who assume that beliefs and desires are *components* of the emotion, I endorse a *causalist* reading of BDTE; that is, I assume that the belief and desire together cause the emotion, which is (accordingly) regarded as a separate mental state.[1] Arguments for this position are presented in Reisenzein (2012a).

BDTE does not claim to be able to explain all mental states that may be presystematically subsumed under the category "emotion". However, the theory wants to explain all those emotions that seem to be directed at propositional objects, that is, actual or possible states of affairs. According to my explication of BDTE, all of these "propositional" emotions are reactions to the cognized actual

[1] Specifically, emotions in CBDTE are conceptualized as nonpropositional signals that are subjectively experienced as feelings of, in particular, pleasure and displeasure, surprise and expectancy confirmation, and hope and fear (see the next section, and Reisenzein 2009a).

or potential fulfillment or frustration of desires; plus, in some cases (e.g., relief and disappointment), the confirmation or disconfirmation of beliefs (Reisenzein 2009a). For example, Maria is *happy* that $p$ (e.g., that *Mr. Schroiber was elected chancellor*) if she desires $p$ and now comes to believe firmly (i.e., is certain) that $p$ is the case; whereas Maria is *unhappy* that $p$ if she is averse to $p$ (which is here interpreted as: she desires $\neg p$ [*not-p*]) and now comes to believe firmly that $p$ is the case. Maria *hopes* that $p$ if she desires $p$ but is uncertain about $p$ (i.e., her subjective probability that $p$ is the case is between 0 and 1), and she *fears* $p$ if she is averse to $p$ and is uncertain about $p$. Maria is *surprised* that $p$ if she up to now believed $\neg p$ and now comes to believe $p$; she is *disappointed* that $\neg p$ if she desires $p$ and up to now believed $p$, but now comes to believe $\neg p$; and she is *relieved* that $\neg p$ if she is averse to $p$ and up to now believed $p$, but now comes to believe $\neg p$. The analysis of social emotions is discussed below.

## *1.1 A Computational Model of BDTE*

Like most traditional theories of psychology, including most emotion theories, BDTE is formulated on the "intentional level" of system analysis (in Dennett's 1971, sense) familiar from common-sense psychology; in fact, BDTE is an explication of a core part of the implicit theory of emotion contained in common-sense psychology (Heider 1958). However, I believe with Sloman (1992) that some basic questions of emotion theory can only be answered if one moves beyond the intentional level of system analysis to the "design level", the level of the computational architecture (Reisenzein 2009a, b). This requires making assumptions about the representational-computational system that generates the mental states (beliefs, desires, emotions) assumed in BDTE. The computational architecture that I have adopted as the basis for a computational model of BDTE assumes a propositional representation system, a "language of thought" (Fodor 1975, 1987). The main reason for this architectural choice is that, in contrast to other proposed representation systems (e.g., image-like representations, or subsymbolic distributed representations of the neural network type), a language of thought provides for a plausible and transparent computational analysis of beliefs and desires. In fact, considering that the intentional objects of beliefs and desires are generally regarded as propositions or states of affairs, and that propositions are the entities described by (are "the meanings of") declarative sentences, a propositional representation system seems to be the natural choice for the computational modeling of beliefs and desires. If one combines this assumption about the representational format of the contents of beliefs and desires with the basic postulate of cognitive science, that mental processes are computations with internal representations, then one immediately obtains Fodor's (1987) thesis that the mental states of *believing* and *desiring* are special modes of processing propositional representations, that is, sentences in the language of thought. To use Fodor's metaphor, believing that a state of affairs $p$ is the case consists, computationally speaking, of having a token of a sentence $s$ that represents $p$ in a special memory store (the "belief store"), whereas desiring *p* consists of having a token of a sentence *s* that represents *p* in another memory store (the "desire store"). For example, prior to Schroiber's election, Maria desired victory for Schroiber in the election but believed that he would not win it. On the computational level, this means that prior to Schroiber's election, Maria's desire store contained among others the sentence "Schroiber will win the election", and her belief store contained the sentence "Schroiber will not win the election."

CBDTE also follows Fodor (1975) in assuming that at least the central part of the language of thought is innate. In particular, CBDTE assumes that the innate components of the language of thought comprise a set of hardwired maintenance and updating mechanisms (Reisenzein 2009a). At the core of these mechanisms are two comparator devices, a *belief-belief comparator* (BBC) and a *belief-desire comparator* (BDC). As will be explained shortly, these comparators play a pivotal role in the generation of emotions. The BBC compares newly acquired beliefs to pre-existing beliefs, whereas the BDC compares them to pre-existing desires. Computationally speaking, using again Fodor's "store" metaphor, the BBC and BDC compare the mentalese sentence tokens $s_{new}$ in a special store reserved for newly acquired beliefs, with the sentences $s_{old}$ currently in the stores for pre-existing beliefs and desires. If either a match ($s_{new}$ is identical to $s_{old}$) or a mismatch ($s_{new}$ is identical to $\neg s_{old}$) is detected, the comparators generate an output that signals the detection of the match or mismatch.

CBDTE assumes that the comparator mechanisms operate automatically (i.e., without intention, and preconsciously) and that their outputs are *nonpropositional* and *nonconceptual*: They consist of signals that vary in kind and intensity, but have no internal structure, and hence are analogous to sensations (e.g., of tone or temperature, Wundt 1896). These signals carry information about the degree of (un-) expectedness and (un-) desiredness of the propositional contents of newly acquired beliefs; but they do not represent the contents themselves. In our example, Maria's BBC detects that the sentence $s_{new}$ representing that Schroiber wins the election, is inconsistent with (is the negation of) the content $s_{old}$ of a pre-existing belief; and Maria's BDC detects that $s_{new}$ is identical to the content $s_{old}$ of an existing desire. As a consequence, Maria's BBC outputs information about the detection of a mismatch—the information that one of Maria's beliefs has just been disconfirmed by new information; whereas Maria's BDC outputs information about a match—the information that one of Maria's desires has just been fulfilled.

To complete the picture, CBDTE assumes that the outputs generated by the BBC and BDC have important functional consequences in the cognitive system. First, attention is automatically focused on the content of the newly acquired belief that gave rise to match or mismatch—in Maria's case, Schroiber's unexpected but desired election victory. Second, some minimal updating of the belief-desire system takes place automatically: Sentences representing disconfirmed beliefs are deleted from the belief store, and sentences representing states of affairs now believed to obtain are deleted from the desire store. Third, BBC and BDC output signals that exceed a certain threshold of intensity give rise, directly or indirectly, to unique conscious feeling qualities: the feelings of surprise and expectancy confirmation (BBC), and the feelings of pleasure and displeasure (BDC). It is assumed that the simultaneous experience of an emotional feeling and the focusing of attention on the content of the belief that caused it, give rise to the subjective impression that emotions are directed at objects (Reisenzein 2009a).

In sum, CBDTE posits that emotions are the results of computations in a propositional representation system that supports beliefs and desires. The core of the belief-desire system is innate, and this innate core includes a set of hardwired monitoring-and-updating mechanisms, the BBC and the BDC. These mechanisms are, in a sense, similar to sensory transducers (sense organs for color, sound, touch, or bodily changes); in particular, their immediate outputs are nonpropositional and nonconceptual, sensation-like signals. However, instead of sensing the world (at least directly), these "internal transducers" sense the current state and state changes of the belief-desire system, as it deals with new information. Emotions result when the comparator mechanisms detect a match or mismatch between newly acquired beliefs and pre-existing beliefs (BBC) or desires (BDC). Hence, according to CBDTE, emotions are intimately related to the updating of the belief-desire system. In fact, the connection could not be tighter: The hardwired comparator mechanisms that service the belief-desire system, the BBC and the BDC, *are simultaneously the basic emotion-producing mechanisms*. Correspondingly, CBDTE assumes that the evolutionary function of the emotion mechanisms is *not* to solve domain-specific problems (as proposed by some evolutionary emotion theorists; e.g., Ekman 1992; McDougall 1908/1960; Tooby and Cosmides 1990), but the domain-general task to detect matches and mismatches of newly acquired beliefs with existing beliefs and desires, and to prepare the cognitive system (or agent) to deal with them in a flexible, intelligent way once they have been detected.

As explained in more detail in Reisenzein (2009a, b), CBDTE solves, resolves, or at least gives clear answers to several long-standing controversial questions of emotion theory. For example, CBDTE provides a precise theoretical definition of emotions (Reisenzein 2009b, 2012a): Emotions are the nonpropositional signals generated by the belief- and desire congruence detectors, that are subjectively experienced as unique kinds of feeling. In contrast to other evolutionary emotion theories, CBDTE also provides a principled demarcation of the set of basic emotions: This set includes precisely the different kinds of output of the congruence detectors. At the same time, however, CBDTE speaks against a sharp distinction between "basic" and "nonbasic" emotions: All emotions covered by the theory, however complex or culturally determined they might be in other respects (this concerns in particular the social emotions), are equally basic in the sense that they are all products of innate, hardwired emotion mechanisms, the BBC and the BDC. CBDTE also provides an explanation of the phenomenal quality of emotions—the fact that emotional experiences "feel in a particular way" (Reisenzein 2009b; see also, Reisenzein and Döring 2009)—and it is able to give a plausible picture of the relation of emotions to public language (Reisenzein and Junge 2012). Finally, an extension of CBDTE to "fantasy emotions"—emotions elicited by stage plays, novels, films etc., as well as by the vivid imagination of events—has been proposed in Reisenzein (2012b, c).

## 2 Social Emotions from the Perspective of CBDTE

Like BDTE, CBDTE assumes that most emotions are variants of a few basic forms. Accordingly, the apparent diversity and complexity of emotional experiences is not due to the existence of many different "discrete" emotion mechanisms, as some emotion theorists have claimed (e.g., McDougall 1908; Ekman 1992; Tooby and Cosmides 1990). Rather, it is due to the fact that humans can have complex beliefs and desires. Specifically, as members of an "ultrasocial" species (Richerson and Boyd 1998), humans have a vital interest in the fate of (certain) others, their actions and mental states, and the effects of their own actions on others (see also, Reisenzein and Junge 2012). That is, these "social objects" are preferred objects of the beliefs and desires of humans. And if these social beliefs and desires are processed by the emotion mechanisms postulated in CBDTE, social emotions may result. The truth of this assumption can only be established by demonstration, that is, by providing analyses of specific social emotions from the perspective of CBDTE. Here, I will analyze pity as the representative of the fortune-of-others emotions, and guilt as the representative of the norm-based emotions; however, the results of the analyses can be generalized, with small adaptations, to other fortunes-of-others emotions and other norm-based emotions respectively. The following analyses are based, on the one hand, on CBDTE, which provides the theoretical framework for the analysis; and on the other hand, on my intuitions about the social emotions.[2]

### 2.1 *Emotional Reactions to the Fate of Others: The Example of Pity*

If CBDTE is true, emotional reactions to the positive and negative fate of other people arise, in principle, in the same way as the emotions of self-regarding happiness and unhappiness about a state of affairs. For the case of pity, this means: Pity is a form of unhappiness—one can also say, a feeling of suffering, or a kind of displeasure or mental pain (Miceli and Castelfranchi 1997)—that occurs, like all "propositional" emotions of displeasure, if the belief-desire comparator (BDC) discovers that the content of a newly acquired belief contradicts the content of an existing desire. As an example, imagine that Maria learns that her colleague *Karl* (= *o*) *has lost his job* (= *F*), and that Maria experiences pity with Karl because of this

[2] I consider these intuitions to be the results of mental simulations of situations that elicit social emotions, and therefore accord them empirical status, although my simulations are limited by being single-case experiments. However, some of the simulation results reported for pity have been replicated in larger samples using hypothetical scenarios (Reisenzein 2002). In any case, readers are invited to join in the described mental simulations.

state of affairs *Fo*.[3] According to CBDTE, the proximate causes of Maria's pity with Karl are her belief that Karl lost his job, *Bel*(*Fo*), and her desire that Karl should not lose his job, *Des*($\neg$*Fo*). Given these inputs, Maria's BDC detects that the content of a newly acquired belief (*Fo*) contradicts the content of one of her desires ($\neg$*Fo*) and as a consequence generates a nonpropositional, sensation-like signal that represents the detection of this desire-incongruence, and that is experienced by Maria as a feeling of displeasure or mental pain.[4]

As presented so far, the only difference between pity (for another) and self-regarding unhappiness is that the desire that is frustrated in pity concerns the fate of another social agent. However, this analysis is clearly insufficient to individuate pity as a separate emotion, as a distinct form of unhappiness or emotional suffering. In fact, at second sight this analysis does not even allow to distinguish pity from self-regarding unhappiness. Imagine, for example, that Maria suffers from Karl's job loss solely because she believes that it will cause her own work situation to deteriorate (she believes that she will have to take over part of Karl's work), but that apart from this, Maria is completely indifferent to Karl's fate. In this case—my intuition tells me—Maria will be unhappy *that Karl lost his job*, but she will not be unhappy *with Karl* about the loss of his job; or to put it differently, she will feel sorry *because of* Karl's job loss, and will feel sorry *for herself* about his job loss, but she will not feel sorry *for Karl* because of his job loss—she will not feel pity for Karl (for empirical evidence, see Reisenzein 2002).[5] Hence, believing that an undesired event has occurred that affects another person is insufficient for experiencing pity for the other. To experience pity for Karl, more is needed on Maria's part than her desire that Karl is not fired from his job, and her belief that that he has been fired.

Now, whatever this additional factor is, if CBDTE is correct, it cannot be another *proximate* cause of Maria's pity. The reason is that, according to CBDTE, being displeased about a state of affairs *p* has exactly *two* proximate causes, the belief that *p*, and the desire that $\neg p$. Only these two mental representations are (direct) inputs to the BDC, the mechanism that produces hedonic feelings. Therefore, the problem posed by pity for CBDTE—how to individuate pity as a special form of

[3] In this example, the object of pity is a social event involving the other (Karl's job loss). However, in general the object of pity can be any state of affairs involving another agent: His social or physical condition, his mental states (beliefs, desires, emotions), and his actions.

[4] According to CBDTE, this feeling, like all emotions, is in itself not object-directed; it has no propositional content (Reisenzein 2009a). However, CBDTE assumes that, as a result of being processed by subsequent cognitive processes, the feelings generated by the emotion mechanisms are usually linked to the propositional objects of their causative beliefs, giving rise to the subjective impression that the feelings are directed at these objects (see Reisenzein 2009a, 2012a). In our example, Maria's feeling of displeasure is linked to *Fo*; as a result, it appears to Maria that she feels sorry about *Fo*. Presupposing this understanding of object-directedness, it is unproblematic to speak about the intentional object of pity and other emotions in CBDTE.

[5] The existence of a special grammatical construction (the *feeling-for* construction) in ordinary language to describe other-regarding emotions (e.g., *feeling sorry for*, *fearing for*, *hoping for*, *being angry for* someone) indicates that the distinction between self- and other-regarding feelings is salient and important in common-sense psychology.

suffering—cannot be solved by assuming that pity has another (direct) mental cause. One could, of course, decide to modify CBDTE to allow this to be the case; but this would mean to give up a basic and (I believe) intuitively compelling idea on which CBDTE is founded, namely the assumption that all (hedonic) emotions result from a match or mismatch between what one believes to be the case, and what one desires. Furthermore, it is not clear what the additional cause of pity—the third input to the BDC—could be.

However, there is another solution: Pity and self-regarding unhappiness could differ in terms of their cognitive-motivational "background"; specifically, they could differ in terms of the beliefs and desires on which the proximate desire for another's fate is based. In other words, the difference between Maria's pity with Karl because of its job loss, and her self-regarding sorrow could reside in the *grounds* or *reasons* for which Maria finds Karl's job loss undesirable.

### 2.1.1 Pity Has a Special Cognitive-Motivational Background

Most desired states of affairs are not desired for their own sake, but because they are believed to lead to other, desired states of affairs, or at least to increase their likelihood of occurrence. In other words, most desires are derived from other, more basic desires. Typically, the derivation of desires form others is achieved with the help of means-ends beliefs (Conte and Castelfranchi 1995; Reisenzein 2006b; Reiss 2004): One desires $p_1$ because one desires $p_2$ and believes that $p_1$ will lead to $p_2$; one desires $p_2$ because one desires $p_3$ and believes that $p_2$ will lead to $p_3$; and so on, until one eventually arrives at a state of affairs that one no longer desires as a means to another end, but for its own sake.[6] Desires for such states of affairs are *basic motives*. Plausible candidates for basic motives are biological urges (the desire for food, sex, physical integrity, etc.); but many "higher" motives of humans, such as the hedonistic motive, the power motive, the affiliation motive, or the desire for knowledge, are also regarded as basic motives by some authors (see Reisenzein 2006b; Reiss 2004). Of particular importance for the present analysis, according to a school of motivation psychology that dates back to, at least, Adam Smith (1759), if not to Aristotle,[7] humans have not only egoistic but also altruistic motives: desires for the well-being of (suitable) other people that are not derived from egoistic motives (see Batson and Shaw 1991; Sober and Wilson 1998). That is, we sometimes desire the well-being of others for their own sake, and not because we believe to profit from it. Empirical evidence for the existence of altruistic motives

[6] The derivation of desires from other desires can occur by means of conscious reasoning processes, that is by reflecting about one's desires and possible means to satisfy them; however, the most basic mechanism of desire-derivation consists presumably of a hardwired procedure that automatically generates a derived desire for $p_1$ whenever a more basic desire $Des(p_2)$ and a fitting means-ends belief $Bel(p_1 \rightarrow p_2)$ are present (see also, Conte and Castelfranchi 1995).

[7] In the *Rhetoric*, Aristotle defines friendship as wanting good things for another for his sake and not for one's own. See Cooper (1977).

has been provided, in particular, by the social psychologist C. Daniel Batson and his co-workers (e.g., Batson and Shaw 1991). In the following, I will accept the altruism hypothesis as correct. Interestingly, however, the existence of altruistic motives is *independently* suggested by the analysis of pity proposed here. That is, this analysis suggests that, to explain pity (and other fortune-of-others emotions), specifically to distinguish pity from self-regarding unhappiness, it is necessary to assume that humans have altruistic motives. Hence, the CBDTE analysis of pity provides an independent reason for believing in the existence of altruistic motives (see also Reisenzein 2002).

Maria's desire that Karl should not lose his job, and similar concrete desires concerning the fate of other people, are certainly not basic motives but are derived from other desires.[8] This opens up the possibility that pity for another and self-regarding sorrow evoked by another's fate are based on different background desires. That this is indeed the case, is in fact rather directly suggested by a closer inspection of our example case: If Maria feels only self-regarding sorrow when she learns about Karl's job loss, then she presumably wants Karl to keep his job only because she believes that Karl's job loss will harm her own well-being; or in short, that his job loss is bad *for her*. In contrast, if Maria feels pity for Karl because he lost his job, she presumably desires Karl to keep his job, at least in part, because she believes that the job loss will harm Karl's welfare, or for short, that it is bad *for Karl*.

However, Maria's belief that Karl's job loss will have negative consequences for him is clearly not by itself sufficient to derive Maria's desire that Karl should keep his job. In addition to this means-ends belief, Maria must also have another, more basic desire from which the former desire can be derived. The obvious candidate for this background desire is Maria's desire that good things should happen to Karl and that he be spared bad things, at least within reasonable limits (e.g., that Karl gets what he deserves; see Ortony et al. 1988). By contrast, in the case of Maria's self-regarding sorrow about Karl's job loss, her desire that Karl should keep his job is derived from her desire to avoid the negative consequences that Karl's job loss would have for her (e.g., the extra work she would have to do), together with her belief that this goal will be reached if Karl stays employed.

However, the CBDTE analysis of pity is still not complete. Rather, one must ask further where Maria's desire for Karl's welfare stems from. My thesis is: If Maria is to experience pity with Karl about his job loss, rather than just self-regarding unhappiness, then her desire for Karl's welfare must not be derived (exclusively) from egoistic desires. Rather, her desire for Karl's welfare must be (at least in part) altruistic. This thesis can be supported by both theoretical and empirical arguments.

[8] The derivation of the concrete desire that proximately causes pity often occurs only at the time when pity is experienced, because this derivation is often occasioned by becoming aware of the event that elicits pity. However, the derivation of this desire can of course take place earlier, analogous to the case of Maria's happiness about Schroiber's election victory. For example, Maria could have formed the (explicit) desire that Karl should not lose his job when she heard about upcoming personnel cuts. When she later heard about Karl's job loss, that desire only had to be retrieved from long-term-memory.

First the theoretical arguments. If Maria desires Karl's general well-being only for egoistic reasons (e.g. because she thinks that, as long as Karl is doing well, he will be an asset rather than a burden) then her concrete desire that Karl should keep his job, which is derived from the former desire, is also egoistic—one cannot derive an altruistic desire from egoistic motives. However, in this case, Maria is in the same kind of motivational situation as when she hopes to profit directly and concretely from Karl's continued employment (i.e., because she does not have to take over additional work), as discussed earlier. Karl's job loss should therefore again only cause Maria to feel self-regarding unhappiness, but not pity for Karl. Furthermore, as mentioned above, the desire for the well-being of another person is just the desire that the other should, within reasonable limits, experience good things and should be spared bad things. As argued above, however, the frustration of a single, concrete desire of this kind (e.g., that Karl should keep his job) does not evoke pity if it is purely egoistically motivated. If so, it is hard to see how a desire for many, more abstractly described events of the same kind ("Karl should be spared negative events") could form the motivational basis of pity.

The empirical support for the claim that, to experience pity for another, one must not desire the other's welfare for egoistic reasons only, consists of the results of mental simulations of diagnostic hypothetical situations. How would one react emotionally if something negative happens to another person whose welfare one desires only for selfish reasons? As an extreme case, one might imagine a slave who loathes his cruel master but is nonetheless concerned about his well-being because his fate depends completely on that of the master: Any deterioration of the master's welfare will immediately be felt by the slave. What emotions will this slave experience when he learns that his master has, say, suffered a severe economic loss? Probably self-regarding sorrow, and fear; but according to my intuition, not pity.

However, if Maria's desire for Karl' welfare is not derived from egoistic motives, then this desire is either itself a basic motive—that would be a person-specific, basic altruistic motive (i.e., one that concerns a specific person, Karl)—or it is derived from more basic nonegoistic motives. Specifically, Maria's desire for Karl's welfare could be derived from her basic altruistic desire that suitable people (such as friends) should experience, within reasonable limits, good things and should be spared bad things.

The proposed CBDTE analysis of pity can be summarized as follows.

**CBDTE analysis of pity**: Pity about $p$ is a form of unhappiness (suffering, mental pain) about $p$, that person $a$ experiences if:

1a. $a$ believes that $p$; with $p = Fo =$ a state of affairs of type $F$ that concerns another person (or group) $o$; and

2b. $a$ desires that $\neg Fo$.

3. $a$'s desire for $\neg Fo$ is based on:

   3a. $a$'s belief that $Fo$ is bad for $o$ (or that $\neg Fo$ is good for $o$) and

   3b. $a$'s desire that (within reasonable limits), good things and no bad things should happen to $o$.

4. The desire 3b (that good things and no bad things should happen to *o*) is not derived from egoistic reasons, but is altruistic.

Because *a*'s desire for the welfare of the other (3b) is altruistic, the desire for $\neg Fo$ that directly causes the emotion (2b) is altruistic as well. Therefore, the proposed analysis of pity can be abbreviated as follows: Pity about *p* is a form of unhappiness about *p* that is caused by the perception (more precisely: the detection by the BDC) that an altruistic desire has been frustrated by *p*. In contrast, if the desire frustrated by *p* is egoistic, then one experiences self-regarding unhappiness, or egoistic sorrow. Of course, it is also possible that the desire frustrated by *p* is partly derived from altruistic and partly from egoistic motives; indeed, this may be the typical situation. For example, Maria could desire Karl's continued employment both because Karl's well-being is dear to her heart for altruistic reasons, and because she hopes to profit from Karl's continued employment. In this case, Maria experiences a mixture of pity and egoistic sorrow (for evidence, see Reisenzein 2002).

### 2.1.2 Pity Has a Special Intentional Object

There is a second possibility of analyzing pity in CBDTE. The basic idea of this second approach is that the *intentional objects of pity*—the states of affairs that one feels pity about—are of a special kind, a kind specific for pity. The more general intuition behind this analytic approach is that, although the different instances of a given emotion type (happiness, unhappiness, pity, etc.) have different particular objects (e.g., in the case of pity: Karl has lost his job; Berta's marriage proposal was rejected; the kitten caught its paw in the trap), all particular objects of a given emotion have something in common, that can therefore be used to individuate the emotion—it is an emotion whose particular objects share this common property. More formally, for each emotion type $E$ there is a property $P_E$ such that, for all particular objects $p \in \{p_1, p_2, \ldots, p_\mathrm{n}\}$ of $E$, it is true that $P_E(p)$. Philosophers call this common property $P_E$ of the objects of an emotion $E$ the "formal object" of the emotion $E$ (e.g., Kenny 1963; de Sousa 1987; Teroni 2007). As it turns out, the formal object of an emotion is intimately linked to the emotion's cognitive and motivational presuppositions, for the features used to define the formal object are precisely the person's beliefs and desires characteristic for this emotion. Therefore, given any proposed analysis of the beliefs (or the beliefs and desires) characteristic for an emotion, a corresponding formal-object analysis of this emotion falls out as a byproduct. For example, according to BDTE, happiness about *p* is experienced if one desires *p* and comes to believe that *p* is true. Therefore, the formal object of happiness can be described as "the believed occurrence of a desired state of affairs", or in abbreviated form, "the occurrence of a desire-fulfillment"; for all particular objects of happiness, all the things people are (and in fact, can be) happy about, have in common that they consist of the realization of a state of affairs that fulfills a desire of the experiencing person.[9]

What, then, is the formal object of pity? To recall, pity was analyzed as a form of displeasure caused by the belief that a state of affairs *p* obtains which is negative for another person and is undesired for altruistic reasons. This entails that all concrete objects of pity—all particular states of affairs that are and can be objects of pity, regardless of whether they consist of another's loss of job, illness, lovesickness or whatever—have in common that they are, from the cognitive-motivational perspective of the emotion experiencer, present, negative, altruistically undesired states of affairs. To distill the formal object $P_E$ of an emotion *E* from the belief-desire analysis of *E*, one abstracts from the concrete objects of the emotion and characterizes them purely relationally, by referring to the emotion-defining beliefs and desires in which they figure. In this way, the description of the cognitive-motivational basis of an emotion can be packed into a description of the emotion's object, which if nothing else allows to present the results of the belief-desire analysis of emotions in a succinct way (e.g., Lazarus 1991; Ortony et al. 1988; Reisenzein et al. 2003; Roberts 2003). Specifically, making use of the formal object of pity, pity for *o* because of *p* can be described as: unhappiness about a present state of affairs *p* that is negative for *o* and is altruistically undesired.

Because the analysis of pity by means of specifying a formal object is entailed by the belief-desire analysis, it adds at first sight nothing substantial to this analysis. A substantive difference between the belief-desire analysis of emotions and their analysis in terms of corresponding formal objects *would* arise, however, if one assumed that the formal object $P_E$ of an emotion *E* figures not only in the scientists' description of the emotion, but also—in the form of concrete realizations $P_E(p)$—as the intentional object of the person's emotional experience of *E*; or at least as the object of a belief presupposed by *E*. This would mean that to experience pity for Karl, Maria must *herself* represent Karl's job loss as a "present state of affairs that is negative for another, and is altruistically undesired." According to CBDTE, this "formal object cognition" is *not* required to experience pity; certainly not as an explicit belief. One can say, however, that *part of* this cognition—the belief that Karl's job loss is a present event that is negative for Karl—is *implicitly* present; for Maria believes indeed that Karl lost his job, and that Karl's job loss is bad for him. However, the *remaining part* of the formal object cognition, the belief that Karl's job loss is undesirable for Maria for altruistic reasons, need not be present for pity to occur according to CBDTE—neither explicitly nor implicitly. In contrast, the cognitive-evaluative theory of emotion (appraisal theory) seems to imply that,

[9] Roberts (2003, p. 110) speaks of the "defining proposition" of an emotion. In psychology, the appraisal theorist R. S. Lazarus (1991) has coined the very similar concept of "core relational theme". According to Lazarus, every emotion (happiness, sadness, fear and so on) is characterized by a unique core relational theme, which describes what is common to the different specific events that elicit the respective emotion. For example (and differing somewhat from the belief-desire analysis of happiness proposed here), the core relational theme of happiness is "making reasonable progress toward the realization of a goal" (Lazarus 1991, p. 122).

for an emotion to occur, the complete formal object cognition of that emotion must be present as an explicitly represented (although not necessarily conscious) belief. Inasmuch as this implication of appraisal theory is implausible, this is a good reason to be skeptical about it (see Reisenzein 2009a).

However, because CBDTE places no restrictions on the description of potential emotion-eliciting events in the language of thought, it does not exclude the possibility that an event is represented by the person herself as one that (partially) exemplifies the formal object property $P_E$. Therefore, it is at least theoretically possible that Maria pities Karl (not only) for having lost his job, but (also) for the fact that something bad has occurred to him. According to CBDTE, to experience pity of this second kind, Maria must form the explicit belief that something bad has happened to Karl, as well as the explicit desire that this bad thing should not have occurred to him; for only if these explicit representations are available can the BDC access them and generate a feeling of displeasure. A possible example would be the following case: Maria is informed that "a bad thing has happened to Karl", but does not yet know what the bad thing is. According to CBDTE, Maria should in this case first pity Karl that something bad has happened to him, and then—when she learns that Karl lost his job—pity him again for having lost his job.[10] Usually, however, the process of emotion generation works the other way round: Typically, Maria will first learn that a specific event has occurred (for example, that Karl was fired) and will only afterwards, if at all, construct a more abstract representation of this event, such as "something bad has happened to Karl." However, *if* Maria forms this belief, she should not only feel pity for Karl because he lost his job, but also because something bad happened to him—even though these two feelings of pity are probably difficult to distinguish subjectively.

### 2.1.3 Does Pity Feel Special?

The result of the preceding analyses was that pity has a special cognitive-motivational background and as a consequence, a special formal object. I now turn to the emotion itself, the feeling of displeasure that pity is. The question I wish to discuss is: Does the feeling of pity, in addition to having distinct cognitive and motivational causes and a distinct formal object, also have a special phenomenal character? Does it *feel* a particular way to experience the mental pain that pity constitutes, a way that that differs from the feeling of egoistic unhappiness? I think something can indeed be said for this assumption.

The most direct argument for believing that pity is a qualitatively distinct kind of unpleasantness appeals to the evidence of introspection. This evidence suggests to me that pity does indeed feel different from egoistic suffering: It feels different, for

[10] Maria's intensity of pity may however be reduced in the first case, as she does not know exactly how bad the bad thing is that happened to Karl. Furthermore, because of the epistemic uncertainty present in this case, fear may predominate.

example, to experience *sorrow for Karl* for losing his job, and to feel *self-regarding unhappiness* because of Karl's job loss. However, it could be argued that even if one accepts this intuition, it is not clear that the difference in experiential quality between the two experiences is due to a difference in their hedonic tone. It might instead be due to differences in the phenomenal character of the beliefs and desires that cause these emotions, or of the action tendencies that they typically evoke (this objection presupposes that beliefs, desires and action tendencies do have phenomenal quality, which is controversial; see Reisenzein 2012a for more detail).[11] Or perhaps the special experiential quality of pity is due to other emotions that co-occur with pity, such as a feeling of caring for the other, that is absent in self-regarding unhappiness.

However, a more principled argument can be made. This argument focuses on the question of *how we come to know* that we experience pity about another's fate, rather than self-regarding unhappiness (or some other emotion). To be sure, answering this question does not *require* to assume that the displeasure of pity has a special hedonic quality. Even if the hedonic feeling tones of pity and self-regarding unhappiness were exactly alike, we might still be able to *infer* that our emotion is pity from the context of the emotion—its causes and consequences (see Reisenzein and Junge 2012). However, if the proposed analysis of pity is correct, then this "inferential" account of emotion self-ascription assumes a lot: It assumes that, to know that one experiences pity, one must infer that the displeasure one feels about the negative fate of another was caused by the frustration of an altruistic desire, i.e. a desire for the welfare of the other not derived from egoistic motives. Likewise, to know that one experiences self-regarding sorrow, one must infer that the displeasure one feels was caused by the frustration of an egoistic desire. This surely demands too much: After all, many people (scientist and lay persons alike) are not even sure that altruistic desires exist! At this point, the idea that pity has a distinctive hedonic quality becomes attractive. If evolution thought it important to let humans know how they feel about another's fate, beyond pleasure and pain—is the displeasure one feels when learning about another's negative fate sorrow for the other, or just egoistic distress?—but at the same time had to make do with humans' limited capacity for inference, metacognition, and insight into their motives, a natural solution would have been to arrange for it that altruistic and egoistic desires signal their fulfillment and frustration to consciousness by means of qualitatively distinct feelings of pleasure and pain.

The idea that there are several qualitatively distinct kinds of pleasure and pain is of course not new; it was championed by John Stuart Mill (1871; see also West 2004) and was defended, in the emerging academic psychology of emotion, by Wilhelm Wundt (1896) among others (see Reisenzein 2000). Wundt took the idea to its extremes, arguing, for example, that even the pleasurable feelings elicited by tasting sugar and those elicited by tasting mint are qualitatively distinct. For the

[11] To avoid this objection, while still granting partial correctness to the idea that beliefs and desires contribute to phenomenal quality, one could argue that the same pleasure signal feels different if it occurs in different cognitive-motivational contexts (see also Reisenzein 2012a).

present purposes, a much more moderate version of the pluralist theory of hedonic feelings will do, according to which different qualities of pleasure and displeasure are attached to the fulfillment and frustration of different basic motives (or even just some of them). According to this proposal, altruistic and egoistic motives in particular, give rise to distinct nonpropositional signals when the BDC detects that they have been fulfilled or frustrated. These signals are experienced as qualitatively distinct hedonic feelings; and these distinct feelings of pleasure and displeasure allow the person to distinguish between pity and other altruistic feelings (such as joy for the other, but also fear for the other and hope for the other) on the one hand, and self-regarding sorrow and other egoistic feelings (such as joy for oneself, fear for oneself, and hope for oneself) on the other hand. This distinction does not require awareness of the ultimate motives underlying altruistic versus egoistic feelings. What the person notices, however, is that for example the displeasurable feeling evoked by another's fate comes in two different flavors, sorrow for herself, and sorrow for the other.

## 2.2 The Norm-Based Emotions: The Example of Guilt

Given the preceding, detailed analysis of pity, the analysis of guilt as the representative of the norm-based social emotions can be presented in a more succinct way. As implied by referring to this family of social emotions as *norm-based*, I follow tradition in assuming that they are reactions to perceived norm violations (e.g., guilt, indignation) and norm fulfillments (e.g., moral elevation) (see e.g., Ortony et al. 1988). I thus disagree with those authors who have claimed that guilt (and perhaps other norm-based emotions) can be experienced even in the absence of a perceived norm violation (see e.g., Wildt 1993). Although there are cases of guilt that prima facie seem to support this claim, such as the phenomenon of "survivor guilt" (guilt feelings of disaster survivors), I believe that these cases do not withstand scrutiny. Indeed, closer analysis suggests that even in these cases, a norm violation can be found for which experiencers blame themselves (see e.g., Jäger and Bartsch 2006). However, even if the assumption that guilt is *always* caused by a perceived norm violation should prove to be wrong, I take it to be largely undisputed that guilt is so caused *in the standard cases*. Any plausible theory of guilt must be able to explain these standard cases.

If CBDTE is also true for the norm-based emotions, then these emotions are in principle caused in the same way as self-regarding happiness and unhappiness. Specifically, negative norm-based emotions such as guilt are special forms of displeasure, unhappiness, or mental suffering that, like all negative emotions, are caused by the detection of a desire frustration by the BDC mechanism; whereas positive norm-based emotions such as moral satisfaction, like all positive emotions, arise if the BDC detects a desire fulfillment. As an example, let us assume that *Maria* ($= a$) *has lied to her friend Berta* ($= A$) and Maria now feels guilty about her action, the state of affairs $Aa$. According to CBDTE, the proximate causes of Maria's guilt about *Aa* are her belief that she has lied to Berta (*Aa*), and her desire not to have lied to her ($\neg Aa$). Maria's BDC detects that the content of a newly acquired belief (*Aa*) is contrary to the content of one of Maria's desires ($\neg Aa$). As a consequence, the BDC generates a signal that represents the detected desire-incongruence, and that is experienced as a feeling of displeasure or mental pain.

However, analogous to the case of pity, this analysis is insufficient to individuate guilt as a distinct emotion, a separate form of mental suffering. Imagine that Maria is unhappy about having lied to Berta only because her lie turned out to have unfavorable consequences for her (she has brought herself in all sorts of predicaments with it), although Maria has no moral scruples whatsoever about having lied to Berta. In this case—my intuition tells me—Maria will *regret* (feel self-regarding sorrow) that she has lied to Berta; but she will not feel *guilty* about her action.

What, then, is special about the feeling of guilt; what is special about the unhappiness or mental suffering that it is? My proposal is analogous to the case of pity: What is unique about guilt is (at minimum) guilt's special cognitive and motivational background. Specifically, I assume that (1) as highly social creatures (Richerson and Boyd 1998), humans also have desires and beliefs concerning the compliance of others, and themselves, with social and moral norms; and (2) guilt is experienced if one comes to believe that one has done something that conflicts with a behavioral norm, or rule of conduct, that one desires to obey for nonegoistic reasons.

Instead of developing the analysis of guilt step by step, as in the case of pity, I will present the result of this analysis first and comment on it afterwards.

**CBDTE analysis of guilt:** Guilt about *p* is a form of displeasure (or suffering, mental pain) that person *a* experiences if:

1a. *a* believes that *p*; with $p = Aa$, where *Aa* is the performance of an action of type *A* by *a*; and
1b. *a* desires $\neg Aa$ (that s/he had not performed the action *A*).

2. *a*'s the desire for $\neg Aa$ is based on:
   2a. *a*'s belief that *a* is an actor of type *T* and *a* is in situation of type *S*; and
   2b. *a*'s desire that in situations of type *S*, actors of type *T* do not perform actions of type *A*.

3. The desire 2b (the desire for rule compliance) is based on:
   3a. *a*'s belief that in situations of type *S*, actions of type *A* are forbidden for actors of type *T* by a norm-setting agent *P*; and
   3b. *a*'s desire that the commandments and prohibitions of *P* (in general, or at least in this specific case) be obeyed.

4. The desire 3b (that the commandments and prohibitions of *P* be obeyed) is not based on egoistic motives.

In our example, therefore, Maria's desire to not lie to Bertha (1b) is derived from her desire to obey a particular behavioral norm (2b). The contents of norms can generally be described as behavior rules of the form “In situation of type *S*, actors of type *T* should (not) perform actions of type *A*” (e.g., Siegwart 2010). In our example, the relevant rule of conduct can be formulated as “one ought not lie to a friend without need”; or in more detail: “if one communicates with another person who is a friend and there is no important reason for lying, then one should not lie to the other”. From Maria’s desire (2b) that this rule should be adhered to (that actors of type *T* do not perform actions of type *A* in situations of type *S*) and her belief (2a) that she is in a situation of type *S* (a communication situation in which there is no important reason to lie) and that she is an actor of type *T* (she is a friend of Berta, who communicates something to her), one can derive Maria’s specific desire not to lie to Berta in this situation (1b).

One must ask further, however, where Maria’s desire to obey the truth-telling norm (2b) stems from. I propose that Maria’s desire to respect this norm is derived from (3a) Maria’s belief that the behavior (not to lie to a friend) was commanded by a norm-setting agent *P*, and (3b) Maria’s desire to obey the commandments of this authority—either in general, or at least in this specific case. The perceived norm-setting agent can be a single person, a group, society, or a superhuman (god) and even an abstract entity (“the world order”).

The proposed derivation of Maria’s desire (2b) from (3a) and (3b) is an explication of the process of *norm internalization* or *norm acceptance*. It corresponds essentially to a proposal by Conte and Castelfranchi (1995) in cognitive science and to similar views endorsed by a number of psychologists (e.g., Ajzen 1985) and sociologists (e.g., Hart 1994). According to this view, to internalize, or accept, a norm requires not only to *cognize* the norm, i.e. to come to believe that it exists (3a); it also requires to “import” the prescribed action rule into one’s motivational system; or in other words, to *acquire the desire* that the rule be followed (2b). According to the proposed explication of norm internalization, this desire is, like other nonbasic desires, created by deriving it from a more fundamental desire; this case, the desire to obey the commandments of a norm-setting agent *P* (3b). For example, the internalization of the commandment “One ought not lie to a friend without need” by Maria occurs as follows: First, Maria comes to believe (3a) that this commandment exists, i.e. that the compliance with the rule of action, “do not lie to a friend without need” is required by some norm-setting agent *P* (e.g., her parents). Because Maria wishes that the commandments of *P* are obeyed (3b), she acquires the derived desire to obey the truth-telling norm (2b). It may be noted that the belief-desire theory of emotions provides additional support for this “motivational” analysis of norm internalization; for according to the theory, a breach or fulfillment of a norm can cause emotional reactions only if the norm has been accepted in this motivational sense.

Baurmann (2010) has proposed to differentiate the desire to abide by a social or moral norm into two components: the wish that *others* follow the behavior rule that is the content of the norm, and the wish that *oneself* follows it. In the CBDTE analysis of the norm-based emotions, this distinction is represented by specifying whether or not the person counts herself to the group of actors for which the action is commanded. To experience guilt, the latter is required (2a); for if Maria does not count herself among the people addressed by the norm "do not perform *A* in *S*", then the desire for $\neg Aa$ cannot be derived from her belief that she has performed *A*, and her desire that the norm is obeyed. Even then, however, Maria should still be emotionally upset (e.g., indignant) about the norm-violating actions *of other actors b* who, in her opinion, are addressed by the norm. The reason is that, for these actors, the necessary derived desire Des($\neg Ab$) can be computed. For example, most citizens desire and expect that their garbage bins are regularly emptied by the professional garbage collectors, but not by themselves. Consequently, they won't feel guilty if they do not empty their garbage bins, but will be angry if the professionals don't.

Finally, I assume—in analogy to the analysis of pity—that the desire 3b (that the commandments of the norm-setting agent *P* be followed) is *not egoistically motivated*. If one wants to obey a commandment of a norm-setting agent only out of selfish interests, for example to gain reputation or to avoid sanctions, then—my intuition tells me—one may experience *regret* about having performed a norm-violating action; but one will not feel *guilty* about it. The desire to follow the commandment of the norm-setting agent could be derived from *altruistic* desires, or it could be based on the adoption of a group or "we" perspective (Bacharach 2006; Tuomela 2000), resulting in a desire for the welfare of one's group, or "us", which I take to be not entirely egoistic also (since the group includes others in addition to oneself). Finally, Maria's desire to obey the truth-telling norm could be based on a disposition (possibly innate) to accept norms that she considers valid in themselves, independent of any specific norm-setting agent (Heider 1958; see also, Conte and Castelfranchi 1995).[12]

As in the case of pity, an alternative analysis of guilt is possible that makes use of the concept of the formal object of guilt. According to this alternative analysis, guilt is displeasure about the violation of a behavior rule that is, at least in part, accepted (i.e., whose observance is desired) for nonegoistic reasons.

The proposed belief-desire analysis of guilt deliberately leaves out two factors that need to be considered in a complete analysis. The first of these concerns the role played by the *effects* of norm-violating actions for the intensity of guilt: Other factors constant, the intensity of guilt about a norm-violating action is typically greater, the more harm it caused to others. For example, Maria will typically feel more guilt about having lied to Berta if her lie has serious negative consequences for Berta, than if it has only mild or no negative consequences. This can be explained by assuming that in the former case, Maria feels strictly speaking not only guilty about having violated one accepted norm ("do not lie") but also about having violated another norm ("do not cause harm to others"). This proposal is supported by the consideration that if Maria does not regard herself as responsible for the harm caused to Berta (e.g., because it was unforeseeable by her), her guilt will be mitigated. This brings me to the second factor neglected in the proposed analysis of guilt,

[12] Such "objective" norms might be understood as norms that the person believes would be commanded for a human society by an ideal (roughly: a fully informed, fully rational, impartial and benevolent) norm-setting agent.

"perceived responsibility" (cf. Lorini and Schwarzentruber 2011; Ortony et al. 1988; Weiner 2006; actually this factor is also important for pity; see Weiner 2006). In my analysis, I assumed implicitly that the agent held himself responsible for the norm-violating action. One way of how perceived responsibility could be explicitly incorporated into the proposed analysis of guilt is the following: Responsibility modifies the degree to which the person sees herself or others as actors who are addressed by a particular norm.

## *2.3 On the Functions of the Social Emotions*

What are the evolutionary functions of the social emotions, their adaptive benefits? Following a common practice in emotion psychology, I distinguish between "organismic" (system-internal) and social-communicative functions of emotions. According to CBDTE, the *organismic function* of emotions *in general* is to signal matches and mismatches between newly acquired beliefs on the one hand, and existing beliefs and desires on the other hand, to other cognitive subsystems, to thereby globally prepare the agent to deal with them in a flexible and intelligent way. The social emotions fit into this picture well, as can again be seen by considering pity and guilt: According to the proposed analysis, pity signals the frustration of an altruistic desire by another's negative fate; whereas guilt signals the frustration, by an own action, of a nonegoistic desire to comply with a norm. I assume it is important to become immediately and distinctly aware of these changes in the belief-desire system whenever they occur. Simultaneous with the experience of the unpleasant feelings, the person's attention is automatically drawn to the emotion-evoking events—the negative fate of the other in the case of pity, and the own deviant action in the case of guilt—thereby enabling and preparing the conscious analysis of these events, their causes and consequences (Reisenzein 2009a).

Pity and guilt typically evoke action tendencies to change the eliciting situations—for example to help the other in the case of pity, or to make an attempt to repair an inflicted damage in the case of guilt. According to a number of authors (e.g., Weiner 2006), these action desires are generated by the respective emotions in a direct, nonhedonistic way (see Reisenzein 1996). This assumption is also adopted in CBDTE. In addition, however, it is assumed that the experience of the unpleasant feelings of pity and guilt may reinforce the person's action motivation, by generating hedonistic motives to reduce the unpleasant feelings (Reisenzein 2009a). In CBDTE, the hedonistic mechanism is hence regarded as a *motivational support* mechanism: a mechanism that reinforces the motivation to satisfy the original desire that *p*, when it is threatened or frustrated, by creating an auxiliary desire to reduce or abolish the displeasure caused by a threat to, or a frustration of, the primary desire. In this way, the secondary, hedonistic desire reinforces the primary desire even though it is, in and of itself, blind to the aim of the primary desire. In addition, specifically in the case of guilt, anticipatory feelings experienced if one vividly imagines a possible rule violation—which presumably engages the emotion mechanisms in a “simulation mode” (Reisenzein 2012b)—can prevent a norm violation from occurring in the first place. The helping and rule-abiding actions caused by pity and guilt, respectively, presumably increase the reproductive chances of individuals or (if one accepts the possibility of group selection; see Richerson and Boyd 2005) those of social groups (see also below).

In addition to their system-internal functions, emotions are frequently assumed to have *social-communicative* functions: adaptive benefits that accrue from the involuntary (and perhaps also the voluntary) signalling of emotions to others. According to CBDTE, paralleling the system-internal function of emotions, the verbal or nonverbal communication of emotions to other agents informs them about the occurrence of a belief-belief or belief-desire match or mismatch in the communicating agent. Thereby, communicated emotions alert others simultaneously to two changes: (a) The agent acquired a new belief that matched or mismatched a pre-existing belief or desire; and (b) something may have occurred in the world that caused this agent to experience a belief-belief or belief-desire match or mismatch (Reisenzein and Junge 2012). It is easy to see how this information could be useful *for other agents*: It allows them to update their mental model of the emotion experiencer, or of the environment, and thereby to better adapt to either.

However, what are the adaptive benefits of communicating social (and other) emotions *for the communicator*? At first sight, there seem to be only disadvantages: By communicating his or her emotions to others, the agent becomes more predictable and thus exploitable by others, and gives away potentially useful information about the environment for free. The readiness to (truthfully) communicate emotions, if it exists at all, is therefore a form of biological altruism. As such, it would have required special evolutionary conditions to emerge. Possible evolutionary scenarios are kin selection, reciprocal altruism, group selection (Richerson and Boyd 2005), and costly signalling (Dessalles 2007). My own bet with respect to the social emotions is the group selection scenario. Even though it may not usually be of advantage *for an individual* to reveal his emotions and thereby the underlying desires to others, I submit the hypothesis that *groups* in which emotions, in particular social emotions, are truthfully communicated, are at an advantage over groups in which emotions are hidden or faked. The signalling of social emotions may therefore have been selected as a truthful sign of other’s *group-centered* concerns: Their altruistic concern for others, and their true caring for the observance of the social norms of the group. In showing pity for Karl about his job loss, Maria reveals to Karl and to others that Karl’s fate is “genuinely” (that is, according to the proposed analysis, not just for selfish reasons) dear to her heart; and in showing guilt about having lied to Berta, Maria reveals to Berta and others that she “truly” cares about the social norm that she violated (that is, she wants to obey the norm not only for selfish reasons), and is therefore a particularly reliable adherent of the group norms. Hence, the social-communicative function of the social emotions is to reveal others’ social (nonegoistic) desires. This function is in my view central to understanding the role that emotions play for the stabilization of social order.

## References

Adam, C., A. Herzig, and D. Longin. 2009. A logical formalization of the OCC theory of emotions. *Synthese* 168: 201–248.

Ajzen, I. 1985. From intentions to actions: A theory of planned behavior. In *Action control: From cognition to behavior*, ed. J. Kuhl and J. Beckmann, 11–39. Berlin: Springer.

Arnold, M.B. 1960. *Emotion and personality*, vol. 1 & 2. New York: Columbia University Press.

Bacharach, M. 2006. In *Beyond individual choice: Teams and frames in game theory*, ed. N. Gold and R. Sudgen. Princeton: Princeton University Press.

Batson, C.D., and L.L. Shaw. 1991. Evidence for altruism: Toward a pluralism of prosocial motives. *Psychological Inquiry* 2: 107–122.

Baurmann, M. 2010. Normativität als soziale Tatsache. H. L. A. Harts Theorie des "internal point of view" [Normativity as a social fact. H. L. A. Hart's theory of the "internal point of view". In *Regel, Norm, Gesetz. Eine interdiziplinäre Bestandsaufnahme [Rule, norm, law. An interdisciplinary survey]*, ed. M. Iorio and R. Reisenzein, 151–177. Frankfurt am Main: Peter Lang Verlag.

Castelfranchi, C., and M. Miceli. 2009. The cognitive-motivational compound of emotional experience. *Emotion Review* 1: 223–231.

Conte, R., and C. Castelfranchi. 1995. *Cognitive and social action*. London: UCL Press.

Cooper, J.M. 1977. Aristotle on the forms of friendship. *Review of Metaphysics* 30: 619–648.

Davis, W. 1981. A theory of happiness. *Philosophical Studies* 39: 305–317.

De Sousa, R. 1987. *The rationality of emotions*. Cambridge: MIT Press.

Dennett, D.C. 1971. Intentional systems. *Journal of Philosophy* 68: 87–106.

Dessalles, J.-L. 2007. *Why we talk: The evolutionary origins of language*. Oxford: Oxford University Press.

Ekman, P. 1992. An argument for basic emotions. *Cognition & Emotion* 6: 169–200.

Ellsworth, P.C., and K.R. Scherer. 2003. Appraisal processes in emotion. In *Handbook of affective sciences*, ed. R.J. Davidson, K.R. Scherer, and H.H. Goldsmith, 572–595. Oxford: Oxford University Press.

Fodor, J.A. 1975. *The language of thought*. New York: Crowell.

Fodor, J.A. 1987. *Psychosemantics: The problem of meaning in the philosophy of mind*. Cambridge: MIT Press.

Frijda, N.H. 1986. *The emotions*. Cambridge: Cambridge University Press.

Goldie, P. 2007. Emotion. *Philosophy Compass* 2: 928–938.

Green, O.H. 1992. *The emotions: A philosophical theory*. Dordrecht: Kluwer.

Hart, H.L.A. 1994. *The concept of law*. Oxford: Clarendon.

Heider, F. 1958. *The psychology of interpersonal relations*. New York: Wiley.

Jäger, C., and A. Bartsch. 2006. Meta-emotions. *Grazer Philosophische Studien* 73: 179–204.

Kenny, A. 1963. *Action, emotion, and will*. London: Routledge and Kegan Paul.

Lazarus, R.S. 1991. *Emotion and adaptation*. New York: Oxford University Press.

Lorini, E., and F. Schwarzentruber. 2011. A logic for reasoning about counterfactual emotions. *Artificial Intelligence* 175: 814–847.

Marks, J. 1982. A theory of emotion. *Philosophical Studies* 42: 227–242.

McDougall, W. 1908/1960. *An introduction to social psychology*. London: Methuen.

Meinong, A. 1894. *Psychologisch-ethische Untersuchungen zur Werth-Theorie* [Psychological-ethical investigations in value theory]. Graz: Leuschner & Lubensky. Reprinted in R. Haller & R. Kindinger (Eds.) (1968). *Alexius Meinong Gesamtausgabe* [Alexius' Meinong's complete works] (Vol 3, pp. 3–244). Graz: Akademische Druck- und Verlagsanstalt.

Miceli, M., and C. Castelfranchi. 1997. Basic principles of psychic suffering: A preliminary account. *Theory & Psychology* 7: 769–798.

Mill, J.S. 1871/2001. In: *Utilitarianism*, ed. R. Crisp. Oxford: Oxford University press.

Nussbaum, M.C. 2001. *Upheavals of thought: The intelligence of emotions*. Cambridge: Cambridge University Press.

Oatley, K. 2009. Communications to self and others: Emotional experience and its skills. *Emotion Review* 1: 206–213.

Ortony, A., G.L. Clore, and A. Collins. 1988. *The cognitive structure of emotions*. Cambridge: Cambridge University Press.

Reisenzein, R. 1996. Emotional action generation. In *Processes of the molar regulation of behavior*, ed. W. Battmann and S. Dutke, 151–165. Lengerich: Pabst Science Publishers.

Reisenzein, R. 2001. Appraisal processes conceptualized from a schema-theoretic perspective: Contributions to a process analysis of emotions. In *Appraisal processes in emotion: Theory, methods, research*, ed. K.R. Scherer, A. Schorr, and T. Johnstone, 187–201. Oxford: Oxford University Press.

Reisenzein, R. 2002. Die kognitiven und motivationalen Grundlagen der Empathie-Emotionen. [Cognitive and motivational foundations of the empathic emotions]. Talk delivered at the 43rd Congress of the German Association of Psychologists (DGPs) in Berlin, 2002.

Reisenzein, R. 2006a. Arnold's theory of emotion in historical perspective. *Cognition and Emotion* 20: 920–951.

Reisenzein, R. 2006b. *Motivation* [Motivation]. In *Handbuch Psychologie* [Handbook of psychology], ed. K. Pawlik, 239–247. Berlin: Springer.

Reisenzein, R. 2009a. Emotions as metarepresentational states of mind: Naturalizing the belief-desire theory of emotion. *Cognitive Systems Research* 10: 6–20.

Reisenzein, R. 2009b. Emotional experience in the computational belief-desire theory of emotion. *Emotion Review* 1: 214–222.

Reisenzein, R. 2010. Moralische Gefühle aus der Sicht der kognitiv-motivationalen Theorie der Emotion [Moral emotions from the perspective of the cognitive-motivational theory of emotion]. In *Regel, Norm, Gesetz. Eine interdiziplinäre Bestandsaufnahme* [Rule, norm, law. An interdisciplinary survey], ed. M. Iorio, and R. Reisenzein, 257–283. Frankfurt am Main: Peter Lang Verlag.

Reisenzein, R. 2012a. What is an emotion in the Belief-Desire Theory of emotion? In *The goals of cognition. Essays in honor of Cristiano Castelfranchi*, ed. F. Paglieri, L. Tummolini, R. Falcone, and M. Miceli, 181–211. London: College Publications.

Reisenzein, R. 2012b. Fantasiegefühle aus der Sicht der kognitiv-motivationalen Theorie der Emotion [Fantasy emotions from the perspective of the cognitive-motivational theory of emotion]. In *Emotionen in Literatur und Film [Emotions in literature and film]*, ed. S. Poppe, 31–63. Würzburg: Königshausen & Neumann.

Reisenzein, R. 2012c. Extending the belief-desire theory of emotions to fantasy emotions. *Proceedings of the ICCM* 2012: 313–314.

Reisenzein, R., and S. Döring. 2009. Ten perspectives on emotional experience: Introduction to the special issue. *Emotion Review* 1: 195–205.

Reisenzein, R., and M. Junge. 2012. Language and emotion from the perspective of the computational belief-desire theory of emotion. In *Dynamicity in emotion concepts* (*Lodz Studies in Language*, 27, 37–59), ed. P.A. Wilson. Frankfurt am Main: Peter Lang.

Reisenzein, R., W.-U. Meyer, and A. Schützwohl. 2003. *Einführung in die Emotionspychologie, Band III: Kognitive Emotionstheorien [Introduction to emotion psychology, Vol 3: Cognitive emotion theories]*. Bern: Huber.

Reisenzein, R., E. Hudlicka, M. Dastani, J. Gratch, E. Lorini, K. Hindriks, and J.-J. Meyer. 2013. Computational modeling of emotion: Towards improving the inter- and intradisciplinary exchange. *IEEE Transactions on Affective Computing* 4: 246–266. doi:http://doi.ieeecomputersociety.org/10.1109/T-AFFC.2013.14

Reiss, S. 2004. Multifaceted nature of intrinsic motivation: The theory of 16 basic desires. *Review of General Psychology* 8: 179–193. New York: Tarcher Putnam.

Richerson, P., and R. Boyd. 1998. The evolution of ultrasociality. In *Indoctrinability, ideology and warfare*, ed. I. Eibl-Eibesfeldt and F.K. Salter, 71–96. New York: Berghahn Books.

Richerson, P.J., and R. Boyd. 2005. *Not by genes alone. How culture transformed human evolution*. Chicago: University of Chicago Press.

Roberts, R.C. 2003. *Emotions: An essay in aid of moral psychology*. Cambridge: Cambridge University Press.

Roseman, I.J. 1979, September. *Cognitive aspects of emotions and emotional behavior.* Paper presented at the 87th annual convention of the APA, New York City.

Scherer, K.R. 2001. Appraisal considered as a process of multilevel sequential checking. In *Appraisal processes in emotion: Theory, methods, research*, ed. K.R. Scherer, A. Schorr, and T. Johnstone, 92–129. Oxford: Oxford University Press.

Searle, J. 1983. *Intentionality*. Cambridge: Cambridge University Press.

Siegwart, G. 2010. *Agent – Situation – Modus – Handlung. Erläuterungen zu den Komponenten von Regeln*. [Agent – situation – mode – action. Comments on the components of rules]. In *Regel, Norm, Gesetz. Eine interdiziplinäre Bestandsaufnahme* [Rule, norm, law. An interdisciplinary survey], ed. M. Iorio, and R. Reisenzein, 23–45. Frankfurt am Main: Peter Lang Verlag.

Sloman, A. 1992. Prolegomena to a theory of communication and affect. In *Communication from an artificial intelligence perspective: Theoretical and applied issues*, ed. A. Ortony, J. Slack, and O. Stock, 229–260. Heidelberg: Springer.

Smith, A. 1759. *The theory of moral sentiments*. London: A. Millar.

Smith, M.A. 1994. *The moral problem*. Oxford: Blackwell.

Sober, E., and D.S. Wilson. 1998. *Onto others. The evolution and psychology of unselfish behavior*. Cambridge, MA: Harvard University Press.

Solomon, R.C. 1976. *The passions*. Garden City: Anchor Press/Doubleday.

Steunebrink, B.R., M. Dastani, and J.-J.Ch. Meyer. 2012. A formal model of emotion triggers: an approach for BDI agents. *Synthese* 185: 83–129.

Teroni, F. 2007. Emotions and formal objects. *Dialectica* 61: 395–415.

Tooby, J., and L. Cosmides. 1990. The past explains the present: Emotional adaptations and the structure of ancestral environments. *Ethology and Sociobiology* 11: 375–424.

Tuomela, R. 2000. *Cooperation: A philosophical study*. Dordrecht: Kluwer.

Weiner, B. 2006. *Social motivation, justice, and the moral emotions: An attributional approach*. Mahwah: Erlbaum.

West, H.R. 2004. *An introduction to Mill's Utilitarian ethics*. Cambridge: Cambridge University Press.

Wildt, A. 1993. Die Moralspezifität von Affekten und der Moralbegriff [The normative specificity of emotions and the concept of morality]. In *Zur Philosophie der Gefühle*, ed. H. Fink-Eitel and G. Lohmann, 188–217. Frankfurt/Main: Suhrkamp.

Wundt, W. 1896. *Grundriss der Psychologie*. Leipzig: Engelmann.

# Reasoning with Normative Systems

**Giovanni Sartor**

**Abstract** The cognitive attitudes and operations involved in dealing with large normative systems are significantly different from those involved in complying with isolated social norms. While isolated norms may be directly applied by the agents endorsing them, this does not happen with regard to large normative systems. In the latter case, the agent must first inquire what the system requires from him (or what it allows him to do), namely, what is obligatory or permitted with regard to the normative system, and thus what would be required for complying with it, under different circumstances. I shall propose an argumentation-based approach for enabling an agent to process such requests, as resulting from a normative system and the existing factual circumstances.

**Keywords** Normative systems • Deontic logic • Legal reasoning • Dynamics of legal systems

## 1 Introduction

Human and artificial agents take into account not only shared social norms, but also complex institutional systems. We are often faced with systems of this kind in our daily life (the legal system, but also the prescriptions of an institutionalised religion, or the regulations of a company, a condominium, a regulated market, a teaching institution, a sociotechnical infrastructure such as an airport or a harbour, etc.). Most norms in such systems are created by norm-creating acts of the regulators of such systems (public or private authorities), and their content is to a large extent

Some parts of this paper have been published in the volume "The Goals of Cognition. Essays in Honor of Cristiano Castelfranchi".

G. Sartor (✉)
Law Faculty – CIRSFID, University of Bologna, Bologna, Italy

European University Institute of Florence, Florence, Italy
e-mail: giovanni.sartor@gmail.com

A. Herzig, E. Lorini (eds.), *The Cognitive Foundations of Group Attitudes and Social Interaction*, Studies in the Philosophy of Sociality 5,
DOI 10.1007/978-3-319-21732-1_9

dependent upon the discretionary choice of the regulator. Moreover such norms regulate very specific and differentiated situations, with which most agents do not have previous acquaintance. Thus the precise content of such norms cannot be derived from shared values and attitudes, nor can be induced from social behaviour. Moreover, an institutional system can contain a huge number of such norms, up to hundreds of thousand, so that it exceeds the storage capacity of the human mind. Such norms are also subject to frequent change and to multiple interpretations, so that even if they could be stored in a single repository, the repository would soon become useless unless continuously updated.

This means that for bounded agents, the way of learning the content of a complex institutional system must differ from the way of learning social norms.

When we learn social norms we permanently store them in our memory, as the content of appropriate normative beliefs and goals, so that they can directly govern our behaviour. On the contrary, we do cannot learn and store in our memory most norms included in large normative systems. We rather possess some ideas about the existence of such a system and the ways to identify its content. When needed, we collect some fragmentary information about the system and combine this information with the relevant facts, both tasks being often delegated to experts. On the basis of this information we can conclude that the system requires us to perform certain actions.

When referring to a large normative system *N* an agent usually does not immediately find an answer to the question "What ought I to do?" (as it usually happens when applying a shared social norm the agent is endorsing). The agent rather needs to asks itself (or the appropriate expert) "What does *N* require from me?," i.e., "What ought I do to according to *N*?" Agents may have to deal with different normative systems and distinguish the requests provided by each one of them.

I will propose an argumentation approach, which takes a normative system and the relevant facts as inputs, in order to deliver such answers.

## 2 Preliminary Notions: Actions, Obligations, Norms

For reasoning with normative systems, we need some basic notions. First, a way of expressing action and obligations is required. For actions I will use the simple *E* operator of Pörn (1977) (on the *E* operator see also Sergot 2001), though other action logics, such as STIT (Belnap et al. 2001), would be appropriate as well.

**Definition 1 (Actions).** Let proposition $\mathsf{E}_j\phi$ describe agent *j*'s action consisting in the production of state of affairs $\phi$, where "$\phi$" is any proposition. Thus $\mathsf{E}_j\phi$ means "*j* brings it about that $\phi$". The non-accomplishment of an action is therefore described by $\neg\mathsf{E}_j\phi$, i.e., *j* does not bring about that *A*.

For simplicity, when an agent brings about its own action, I will not repeat the agent's name in the action's result. Thus, for expressing the idea that *John* smokes

(*John* brings it about that he smokes) rather than writing $\mathsf{E}_{John}Smoke(John)$, I will write $\mathsf{E}_{John}Smoke$.

As an example of an action-proposition, consider the following

$$\mathsf{E}_{John}Damaged(Tom)$$

which means "*John* brings it about that *Tom* is damaged", or more simply "*John* damages *Tom*" while the following

$$\neg\mathsf{E}_{John}Damaged(Tom)$$

means "*John* does not bring it about that *Tom* is damaged", or "*John* does not damage *Tom*". I shall adopt the logic of *E*, which is a classical modal logic (if *A* and *B* are logically equivalent, then $\mathsf{E}_xA \leftrightarrow \mathsf{E}_xB$) including the axiom schema:

$$\mathsf{E}_x\phi \Rightarrow \phi$$

meaning that if the state of affairs $\phi$ is realised though an action, then it is the case that $\phi$. For instance, the fact that *Tom* makes it so that *Ann* suffers damage, obviously entails that *Ann* suffers damage:

$$\mathsf{E}_{Tom}Damaged(Ann) \Rightarrow Damaged(Ann)$$

Besides an action logic *E*, I need a deontic logic to express obligations.

**Definition 2 (Obligations and prohibitions).** Let *O* denote obligation. $O\mathsf{E}_j\phi$ means "it is obligatory that *j* brings it about that $\phi$". Similarly $O\neg\mathsf{E}_j\phi$ means "it is obligatory that *j* does not bring about that $\phi$", or "it is forbidden that *j* brings about that $\phi$".

For instance, the following means "it is obligatory that *John* makes it so that *Tom* is compensated", or more simply, "it is obligatory that *John* compensates *Tom*",

$$O\mathsf{E}_{John}Compensated(Tom)$$

while the following means "it is forbidden that *John* damages *Tom*".

$$O\neg\mathsf{E}_{John}Damages(Tom)$$

As usual, I take permission to be the negation of prohibitions. I will not endorse here a particular deontic logic, since the following considerations may apply to different deontic logics. The reader may assume, for instance, standard deontic logic, which is a normal modal logic including, besides all tautologies of propositional logic, definition **Df P**.$P\phi \leftrightarrow \neg O\neg\phi$, axiom **K**.$O(\phi \Rightarrow \psi) \Rightarrow (O\phi \Rightarrow O\psi)$, axiom **D**.$O\phi \Rightarrow P\phi$ and the necessitation rule, **N** according to which if $\phi$ is a theorem, so is $O\phi$. We may on the other hand distinguish obligations directly established by the norms in the normative system of the system and derived obligation extracted from such norms, e.g., indicating what is necessary for complying with them (van der Torre and Hansen 2008). This too would be consistent with our framework, but we cannot explore it here.

For representing legal contents, we need norms, which can be viewed a kind of defeasible conditional.

**Definition 3 (Norm).** A *norm* has the form

$$A \Rightarrow B$$

where $A$ is the antecedent condition, $B$ the ensuing normative conclusion, and $\Rightarrow$ expresses a defeasible unidirectional connection, according to which antecedent $A$ triggers conclusion $B$. In the norm the antecedent $A$ is a proposition and consequent $B$ is any kind of deontic or constitutive normative qualification.

Thus, a norm $A \Rightarrow B$ captures the unidirectional defeasible connection between an antecedent (possibly empty) fact and the normative consequent that is generated by that fact: normative effect $B$ is triggered when the antecedent condition $A$ holds. We write $A \not\Rightarrow B$ for the statement that the norm's antecedent fails to support the norm's conclusion, so that the norm cannot be applied in valid inferences. Arguments establishing that $A \not\Rightarrow B$ undercut the use of the norm $A \Rightarrow B$ in valid inferences.

Here is an example of two deontic norms, the first stating that it is forbidden to cause damage to others, and the second that who causes damage to another has the obligation to compensate the latter (in the following when obvious I drop the requirement $x \neq y$):

$$x \neq y \Rightarrow O\neg \mathsf{E}_x Damaged(y)$$

$$x \neq y \wedge \mathsf{E}_x Damaged(y) \Rightarrow O\mathsf{E}_x Compensated(y)$$

The following is an example of a constitutive norm, saying that if we injure a person (make so that someone is injured), we cause damage to that person (injuring counts as damaging):

$$\mathsf{E}_x Injured(y) \Rightarrow \mathsf{E}_x Damaged(y)$$

Note that I do not distinguish deontic conditionals and constitutive or counts-as conditionals, since both are modelled as defeasible conditionals (Searle 1995; Jones and Sergot 1996).

I assume an argumentation system as defined in Prakken (2010). Such a system includes two sets of inference rules, strict and defeasible inference rules, which have to be applied to a knowledge base of premises.

Strict inference rules have the form $[\phi_1, \ldots \phi_n] \mapsto \psi$. The conclusion $\psi$ of the strict rule holds without exceptions when all its antecedent conditions $\phi_1, \ldots \phi_n$ hold; therefore the application of the rule to derive $\psi$ cannot be challenged unless at least one antecedent condition in $\phi_1, \ldots \phi_n$ is also challenged.

Defeasible inference rules have the form $[\phi_1, \dots \phi_n] \leadsto \psi$; the conclusion $\psi$ of the defeasible rule holds only presumptively (with the possibility of exceptions) when all its antecedent conditions $\phi_1, \dots \phi_n$ hold; therefore the application of the rule to derive $\psi$ can be challenged also without challenging the antecedent conditions, i.e. by rebutting the rule's conclusion or by undercutting the rule's application.

Rules of both kinds can applied to a knowledge base of premises, these being formulas in a logical language.

Here I shall just introduce the main idea of an argumentation system in an informal way. The model I propose is inspired by Prakken (2010), to which I refer for a detailed presentation, though my account will depart from it in some aspects, to provide a simpler framework.

Premises, i.e., formulas in the underlying logical language $\mathscr{L}$ are basic arguments. Further arguments can constructed by applying inference rules to the conclusions of arguments already available: thus given arguments $A_1, \dots A_n$ with conclusion $\phi_1, \dots \phi_n$, through an inference rule $[\phi_1, \dots \phi_n] \mapsto \psi$ we can obtain argument $B_s = \{A_1, \dots A_n \mapsto \psi\}$, while through an inference rule $[\phi_1, \dots \phi_n] \leadsto \psi$ we can obtain argument $B_d = \{A_1, \dots A_n \leadsto \chi\}$. For instance, given premises $a$ and $a \Rightarrow b$ and inference rule $[\phi, \phi \Rightarrow \psi] \leadsto \psi$, we can construct arguments $A_1 = \{a\}$, $A_2 = \{a \Rightarrow b\}$, and $A_3 = \{A_1, A_2 \leadsto b\}$, i.e., $\{\{a\}, \{a \Rightarrow b\} \leadsto b\}$.

Arguments may be defeated (rebutted or undercut) by counterarguments: rebutting takes place when an argument having a conclusion $\psi$ through a defeasible rule (as its ultimate conclusion, or the conclusion of one of its subarguments) faces a non weaker counterargument having the complementary conclusion $\overline{\psi}$; undercutting takes place when an argument including a defeasible rule $[\phi_1, \dots \phi_n] \leadsto \psi$, having name $r$ (we assume that each rule has an unique name) has a counterargument with conclusion $\neg r$ (the negation of a rule-name being understood as the denial of the rule's applicability). An argument is justified, with regard to a knowledge base, if all of its of its defeaters are overruled, being defeated by further justified arguments.[1]

Here I shall not view neither facts nor norms are inference rules of the argumentation system, but rather as premises for it, i.e., as part of its knowledge base (see Prakken and Sartor 2013). Thus, I shall assume a general pattern for building strict inference rules out of modus ponens entailments, and similarly, a general pattern for building defeasible inference rules and undercutters out of defeasible

[1] Argumentation-based semantics (Dung 1995) provides various ways to identify justified arguments, which is done by building maximal sets (called extensions) of the available arguments. For our purposes we can characterise justified arguments as those belonging to an extension that is constructed as follows. We start with the empty set, and progressively admit those arguments that satisfy both of the following conditions: (a) they do not conflict with arguments already admitted, and (b) all their defeaters are defeated by arguments already admitted. The fix-point of this contraction (the set to which no further arguments can be added that satisfy the conditions above) is the so-called grounded extension of an argumentation framework. The same outcome can also be obtained though a dialogue game (Prakken and Sartor 1996; Prakken 2001).

rules (such as, norms). For my purpose, I do not need to address preferences between rules. Therefore the following characterisation of a normative argumentation system will suffice.

**Definition 4 (Argumentation system).** An argumentation system $S$ is a tuple $N_S = (L, R_s, R_d)$ where

- $\mathscr{L}$ is a logical language (here including, in particular, the constructs for propositional logic, action and deontic logic, and the conditional symbols $\Rightarrow$, and $\not\Rightarrow$).
- $R_s$ is the set of all strict inference rules, including
  - Strict modus ponens inference rules: all rules corresponding to the schema $[\phi, \phi \Rightarrow \psi] \mapsto \psi$ for any $\phi$ and $\psi$ in $\mathscr{L}$;
  - Specification: all rules corresponding to the schema $[\phi] \mapsto \phi[t]$, where $t$ is a substitution of variables in $\phi$ with terms in $\mathscr{L}$;
  - Logical axioms: all rules corresponding to the schema $[] \mapsto \phi$, where $\phi$ is any theorem of propositional logic or other deductive logical systems to be used (here action logic **E**, and standard deontic logic **D**}
- $R_d$ is the set of all defeasible inference rules, including
  - Defeasible modus ponens inference rules: all rules corresponding to the schema $[\phi, \phi \Rightarrow \psi] \rightsquigarrow \psi$ for any $\phi$ and $\psi$ in $\mathscr{L}$.
  - Defeasible undercutting inference rules: all rules corresponding to the schema $[\phi \not\Rightarrow \psi] \rightsquigarrow \neg`[\phi, \phi \Rightarrow \psi] \rightsquigarrow \psi'$ where $`[\phi, \phi \Rightarrow \psi] \rightsquigarrow \psi'$ is the name for the inference rule $[\phi, \phi \Rightarrow \psi] \rightsquigarrow \psi$

We read the name of an inference rule as the assertion that the rule is applicable, and so the negation the name of an inference rule is the assertion that the rule is inapplicable; the defeasible undercutting inference schema says that if a norm fails to support its conclusion, then the inference rule based on that norm is inapplicable.

The logical-axioms inference-schema allows any theorem of the deductive logics being used (e.g. $O\phi \rightarrow P\phi$ from deontic logic) to be introduced in any argument, as an unchallengeable premise. We can now define the idea of a normative knowledge base.

**Definition 5 (Normative Knowledge Base).** A normative knowledge base $K$ is a tuple $(C, N)$, of two sets of premises:

- a set $C$ of contextual circumstances,
- a set $N$ of norms (a normative system)

By contextual circumstances I mean the propositions describing the relevant facts of the case, such as the fact that *John* damages *Tom*, the amount of the damage, whether *John* intended to cause the damage or was careless, etc. This notion of a fact is a relative one, since certain normative rules (the constitutive ones) may establish under what conditions a certain qualification is satisfied, so that an apparently factual qualification becomes a normative outcome. Consider for instance legal rules establishing what counts as negligence in road traffic, or in medical practice. For our purpose however, we do not need to address this issue, since we view as contextual circumstances all relevant true propositions that are not established through the application of the normative system we are considering (i.e., that are not established as being the consequent of a norm whose antecedent condition is satisfied).

Finally we define the notion of an entailment with regard to a normative knowledge base and a normative argumentation system.

**Definition 6 (Defeasible entailment).** We shall say that a normative knowledge base $K = (C, N)$ defeasibly entails $A$,and write $K \mid\!\sim A$, to mean that knowledge base $K$ enables us to construct a justified argument for $A$, using the inference rules in argumentation system $S$.

For instance, given knowledge base $K_1 = (\{a\}, \{a \Rightarrow b\})$, we can construct an undefeated (indeed unattached) argument $A_3$ for $b$. Thus we may say that $K_1 \mid\!\sim b$.

Let us now consider knowledge base

$$K_2 = (\{a, c, d\}, \{a \Rightarrow b, c \Rightarrow \neg b, d \Rightarrow (c \not\Rightarrow \neg b)\})$$

This knowledge base enables the construction of argument $A_3$ for $b$, as above.

$K_2$ also enables the construction of arguments $A_4 = \{c\}$, $A_5 = \{c \Rightarrow \neg b\}$ and $A_6 = \{A_4, A_5 \rightsquigarrow \neg b\}$, the latter being a rebutting counterargument to $A_3$.

However $K_2$ also provides for the construction of arguments $A_6 = \{d\}$, $A_7 = \{d \Rightarrow (c \not\Rightarrow \neg b)\}$, $A_8 = \{A_6, A_7 \rightsquigarrow c \not\Rightarrow \neg b\}$ and $A_9 = \{A_8 \rightsquigarrow \neg`([c, c \Rightarrow \neg b] \rightsquigarrow \neg b)'.\}$ The last arguments undercuts $A_5$, so that $A_3$ is freed from its only attacker and is thus justified.

The following example shows how from a norm and an instance of its antecedent we can defeasibly derive an instance of the conditional's consequent.

$$\{\mathsf{E}_{Tom}Damaged(John), \mathsf{E}_{x}Damaged(y) \Rightarrow O\mathsf{E}_{x}Compensated(y)\} \mid\!\sim$$
$$O\mathsf{E}_{Tom}Compensated(John)$$

To execute this inference we just need to instantiate the pattern $[\phi, \phi \Rightarrow \psi] \rightsquigarrow \psi$ into the defeasible inference rule

$$[\mathsf{E}_{Tom}Damaged(John), \mathsf{E}_{x}Damaged(y)$$
$$\Rightarrow O\mathsf{E}_{x}Compensated(y)] \rightsquigarrow O\mathsf{E}_{x}Compensated(y)$$

and apply it to the fact and rule above.

## 3 Relativised Obligations and Permissions

In addressing compliance we have to connect a normative system $N$ and the factual circumstances $C$ relevant to $N$'s application, in the context of a given argumentation system. Here I am only interested in obligations and institutional facts that are generated by norms in $N$, when applied to facts in $C$. Thus we can assume that $C$ contains (or entails) all factual literals (atomic propositions or negations of them) which are true in the real or hypothetical situation in which the norms have to be applied, without considering how the truth of such literals can be established. For simplicity's sake we can limit $C$ to the factual literals that are relevant to the application of norms in $N$, matching literals in the antecedent of a norm in $N$. When the considered factual circumstances are those that hold in the real world (rather than in a merely possible situation), i.e., they are the truths relevant to the application of $N$ in the case at hand, I shall denote them through the expression $T_N$.

I will now introduce the notion of a relativised obligation, namely, a way of expressing the fact that an obligation holds with regard to a normative system and a set of circumstances. A relativised obligation sentence does not express a norm, but it expresses an assertion about the implications of norms (normative systems) and circumstances (in the terminology of Alchourrón 1969 and Alchourrón and Bulygin 1971 such assertions are called "normative propositions").

**Definition 7 (Relativised sentences and obligations).** We say that any sentence $\phi$ holds relatively to normative system $N$ and circumstances $C$, and write $[\phi]_{C,N}$ iff $(C,N) \mid\!\sim \phi$

$$[\phi]_{C,N} \stackrel{\text{def}}{=} (C,N) \mid\!\sim \phi$$

Let the expression $\mathscr{A}_j$ cover both $\mathsf{E}_jA$ and $\neg\mathsf{E}_jA$ and let $\overline{\mathscr{A}_j}$ denote the complement of $\mathscr{A}_j$ ($\overline{\mathscr{A}_j}$ stands for $\neg\mathsf{E}_j\phi$ if $\mathscr{A}_j = \mathsf{E}_j\phi$; it stands for $\mathsf{E}_j\phi$ if $\mathscr{A}_j = \neg\mathsf{E}_j\phi$). Then we can say that it is obligatory to do (or not to do) an action relatively to a certain set of circumstances and a normative system, if such circumstances and system entail that the action ought (or ought not) to be done.

$$\mathbb{O}_{C,N}\mathscr{A}_x \stackrel{\text{def}}{=} (C,N) \mid\!\sim O\mathscr{A}_x$$

When we are referring to the true relevant circumstances of the real world, denoted as $T_N$, rather than to circumstances of hypothetical situations, we simply write $[\phi]_N$, or $\mathbb{O}_N\mathsf{E}_xA$.

$$[\phi]_N \stackrel{\text{def}}{=} (T_N,N) \mid\!\sim \phi$$
$$\mathbb{O}_N\mathscr{A}_x \stackrel{\text{def}}{=} (T_N,N) \mid\!\sim O\mathscr{A}_x$$

For instance, let us consider the following example, where $N_1$ includes a simplified version of the three norms above, and circumstances $C_1$ are limited to the fact that *John* injured *Tom*:

*Example 1.*

$$\begin{aligned} C_1 &= \{\mathsf{E}_{John} Injured(Tom)\} \\ N_1 &= \{\mathsf{E}_x Injured(y) \Rightarrow \mathsf{E}_x Damaged(y) \\ &\quad O\neg\mathsf{E}_x Damaged(y) \\ &\quad \mathsf{E}_x Damaged(y) \Rightarrow O\mathsf{E}_x Compensated(y)\} \end{aligned}$$

It is easy to see that the following inferences holds on the basis of example (1):

$$(C_1, N_1) \mid\!\sim \mathsf{E}_{John} Damaged(Tom)$$
$$(C_1, N_1) \mid\!\sim O\mathsf{E}_{John} Compensated(Tom)$$

Therefore, we can say that *John* has damaged *Tom* and that it is obligatory that *John* compensates *Tom*, relatively to $N_1$ and $C_1$, i.e., that

$$[\mathsf{E}_{John} Damaged(Tom)]_{N_1,C_1}$$
$$\mathbb{O}_{N_1,C_1} \mathsf{E}_{John} Compensated(Tom)$$

If *John* has really injured *Tom* (and no other relevant circumstances obtain, such as exceptions excluding the application of the norms at issue) we can simply say that, according to $N_1$, *John* has damaged *Tom* and it is obligatory that *John* compensates *Tom* i.e.:

$$[\mathsf{E}_{John} Damaged(Tom)]_{N_1} \wedge \mathbb{O}_{N_1} \mathsf{E}_{John} Compensated(Tom)$$

On the basis of example (1) we can also say that it is obligatory that *John* refrains from damaging *Tom*

$$\mathbb{O}_{N_1} \neg\mathsf{E}_{John} Damaged(Tom)$$

Given that it holds that $[\mathsf{E}_{John} Damaged(Tom)]_{N_1}$ we can conclude that the latter obligation has been violated, on the basis of the following definition.

**Definition 8 (Violation).** An obligation $O\mathsf{E}_x A$ of a normative system $N$ is violated in circumstances $C$ iff $(C, N) \mid\!\sim O\mathsf{E}_x A \wedge \neg\mathsf{E}_x A$, In other words the obligation is violated in $C$, iff both $\mathbb{O}_{C,N}\mathsf{E}_x A$ and $[\neg\mathsf{E}_x A]_{C,N}$ hold.

Here is another small example. The first norm in $N_2$ says that if one is in a public place then one is forbidden to smoke. The second says that places open to the public are (count as) public places.

*Example 2.*

$$\begin{aligned} C_2 &= \{OpenToPublic(LectureRoom), in(John, LectureRoom)\} \\ N_2 &= \{OpenToPublic(y) \Rightarrow PublicPlace(y) \\ &\quad PublicPlace(y) \wedge in(x,y) \Rightarrow O\neg \mathsf{E}_x Smoke\} \end{aligned}$$

We can say then say that according to $N_2$ in circumstances $C_2$ it is obligatory that *John* does not smoke ($\mathbb{O}_{C_2,N_2}\neg\mathsf{E}_{Tom}Smoke$).

Clearly, the language of relativised obligations allows us to say that according to different normative systems different obligations hold. For instance, given that Canon law contains both a universal norm prohibiting the use of contraception and a constitutive rule saying any action meant to make a sex act unfruitful counts as artificial contraception, we can conclude that according to the Canon law a woman, say Ann, is forbidden to take the pill in order to have unfruitful sex acts. Similarly, given that Islamic law contains a norm that prohibits receiving interest on loans of money, we can say that according to Islamic law John is forbidden to receive interest on loans of money.

A notion of relativised permission can be provided that corresponds to the above analysis of obligation. While permissions can be modelled as the negation of prohibitions ($P\mathsf{E}_x A \overset{\text{def}}{=} \neg O \neg \mathsf{E}_x A$), relativised permissions can be defined as follows.

**Definition 9 (Relativised permission).** Let us say that it is permissible relatively to $N$ and $C$ that $x$ does (or not does) an action, iff $N$ and $C$ entail that it is permissible to do (or not to do) that action:

$$\mathbb{P}_{C,N}\mathscr{A}_x \overset{\text{def}}{=} (C,N) \mid\!\sim P\mathscr{A}_x$$

Note that according to this definition, saying that an action $\mathsf{E}_x\phi$ is permissible relatively to normative system $N$ and circumstances $C$ ($\mathbb{P}_{C,N}\mathsf{E}_x\phi$) does not amount to saying that it is not the case that $\mathsf{E}_x\phi$ is forbidden relatively to the same system and circumstances ($\neg\mathbb{O}_{C,N}\neg\mathsf{E}_x\phi$). Proposition $\mathbb{P}_{C,N}\mathsf{E}_x\phi$ is not equivalent to $\neg\mathbb{O}_{C,N}\neg\mathsf{E}_x\phi$, since the former holds when $(C,N)$ entails $P\mathsf{E}_x\phi$, while the latter holds when $(C,N)$ does not entail $O\neg\mathsf{E}_x\phi$ (see Alchourrón 1969; Alchourrón and Bulygin 1971).

## 4 Reasoning with Normative Systems

Let us assume that *Tom* wants to know his position concerning the normative systems $L$ (the law). In particular *Tom* is now wondering whether he should pay income tax on the capital gains he obtained by selling his house. Being committed to comply with the law, but not knowing what the law requires from him, *Tom* asks the tax expert *Ann* for advise. Assume that the *Ann* remembers that there is a rule in the tax code that establishes the requirement to pay income taxes on capital gains, but vaguely remembers that there are exceptions to it. This prompts *Ann* to look for exceptions, and she finds indeed one concerning houses. This exception says (in a simplified form) that capital gains from the sale of houses purchased more than 5 year before the sale and inhabited by the seller are exempted from income tax. Assume that Ann's inquiry has led her to conclude that the legal system $L$ she is considering, for instance Italian law, contains the following relevant norms:

$$\begin{aligned} L \supseteq \{ & SellsHouse(x) \Rightarrow O\mathsf{E}_x PayIncomeTaxOnSale, \\ & BoughtMoreThan5YearsBefore(x) \wedge HasInhabitedHouse(x) \\ & \Rightarrow (SellsHouse(x) \not\Rightarrow O\mathsf{E}_x PayIncomeTaxOnSale)\} \end{aligned} \tag{1}$$

where the second norms in (1) says that under the indicated conditions the first one does not hold (is not applicable).

*Ann* then asks *Tom* whether, at the time of the sale, more that 5 years had elapsed from the *Tom*'s purchase, and whether he has been living in the house. Assume that *Tom* replies positively to the first question and negatively to the second one. Then *Ann* says: "Dear, *Tom*, unfortunately you are legally bound to pay income tax on your gains". In fact, by combining the Italian law $L$ with these factual circumstances (let us assume these circumstances are the only relevant ones), *Ann* can see that the following inference holds:

$$\begin{aligned} (\{SellsHouse(Tom), \neg HasInhabitedHouse(Tom)\}, L) \mathrel{|\!\!\sim} & \\ & O\mathsf{E}_{Tom} PayIncomeTaxOnSale \end{aligned}$$

so that, given that both factual premises are true, she can infer what she tells her client:

$$\mathbb{O}_L \mathsf{E}_{Tom} PayIncomeTaxOnSale$$

If *Tom* asks for an explanation, *Ann* would probably answer by saying that whenever one has not lived in the house one sells, then according to the law one has the obligation to pay income tax:

$$SellsHouse(x) \wedge \neg HasInhabitedSoldHouse(x) \Rightarrow \mathbb{O}_L \mathsf{E}_{(}x) PayIncomeTaxOnSale \tag{2}$$

Note that formula (2) does not express a norm of $L$ (there is no norm in $L$ which has exactly that content, see formula (1)). More generally (2) is no norm at all, but rather is a general conditional statement about $L$, namely the statement that in case that the seller has not inhabited the sold house, then $L$ entails that the seller has to pay taxes on capital gains. Similarly, if *Ann* were contacted by *Tom* before making the sale, she would tell him: "Since you have not inhabited the house, if you seel it you will have to pay income tax on your capital gain".

## 5 Dynamic Normative Systems

Let us now consider how an agent (a legislator) can have the ability to introduce new norms in $N$. For this purpose, we need to assume that $N$ is a dynamic normative system (Kelsen 1967), including meta-norms which determine what new norms will be valid according to $N$.

In the framework we have described above, the idea of such a metanorm can be captured through an additional pattern for defeasible inference rules, which enables the production basic norm set to be expanded by further norms. Thus we obtain what we may call a dynamic extension of the normative argumentation system.

**Definition 10 (Dynamic Normative Argumentation System).** A dynamic normative argumentation systems $DS$ is obtained by adding to an argumentation systems $S$ the following inference rules

- Norm creation: all inference rules corresponding to the schema $\mathit{Valid}(\phi) \mapsto \phi$, for any norm $\phi$ in $\mathscr{L}$.

Note that I prefer to model this principle as a strict rule, but depending on how we understand the notion of validity, we could also model it as a defeasible rule (see Sartor 2008).

In a dynamic normative argumentation system, arguments may use rules that do not belong to an initial knowledge base, but that are qualified as valid by norms in that knowledge base. So, let us assume that the knowledge base $K$ of a normative system includes a meta-norm saying that whatever norm $\phi$ is issued by the legislator $\mathit{Leg}$ than $\phi$ is valid (for simplicity's sake I do not consider the temporal dimension of validity, see Governatori and Rotolo 2010).

$$\mathsf{E}_{\mathit{Leg}}\mathit{Issued}(\phi) \Rightarrow \mathit{Valid}(\phi)$$

Given this background, let us assume that legislator accomplished the action of issuing a new norm, for instance, a norm prohibiting any agent $x$ to smoke:

$$\mathsf{E}_{\mathit{Leg}}\mathit{Issued}(O\neg\mathsf{E}_x\mathit{Smoke})$$

The accomplishment of the action described in this formula is a new fact, which is added to the true factual circumstances $T_N$. Thus we can build the following sequence of arguments

- $A_1 = \mathsf{E}_{\mathit{Leg}}\mathit{Issued}(O\neg\mathsf{E}_x\mathit{Smoke})$, premise;
- $A_2 = \mathsf{E}_{\mathit{Leg}}\mathit{Issued}(\phi) \Rightarrow \mathit{Valid}(\phi)$, premise;
- $A_3 = A_2 \mapsto (\mathsf{E}_{\mathit{Leg}}\mathit{Issued}(O\neg\mathsf{E}_x\mathit{Smoke}) \Rightarrow \mathit{Valid}(O\neg\mathsf{E}_x\mathit{Smoke}))$, by specification;
- $A_4 = A_1, A_3 \rightsquigarrow \mathit{Valid}(O\neg\mathsf{E}_x\mathit{Smoke})$, by defeasible modus ponens;
- $A_5 = A4 \mapsto O\neg\mathsf{E}_x\mathit{Smoke}$, by norm creation;
- $A_6 = A5 \mapsto O\neg\mathsf{E}_{\mathit{Tom}}\mathit{Smoke}$, by specification.

Thus, on the basis of argument $A_6$ (which we assume to be unchallenged) we can conclude that smoking is forbidden to *Tom* according to *N*:

$$\mathbb{O}_N \neg \mathsf{E}_{Tom} Smoke$$

## 6 Conclusion

In this paper I have shown how a reasoner may approach a normative system, namely a distinct set of norms, viewing it as an object that enables the derivation of normative conclusions that are relative to that system. For this purpose I have first considered how to model actions, obligations and norms. Then I have defined an argumentation system which takes as inputs knowledge bases of facts and norms, and produces appropriate arguments. On this basis I have considered how obligations and permission can be relative to a particular normative systems, and I have provided a meta-logical representation of this idea. Finally I have developed some considerations on how to model dynamic normative systems in this framework.

While this work is still very preliminary, I hope it can provide some clues on how to model metalevel reasoning with normative systems. Obvious extensions, to be considered in future work, concern integrating this idea with the decision-making process of the concerned agents (for a preliminary attempt, see Sartor 2012), and modelling reasoning with multiple distinct normative systems.

## References

Alchourrón, C.E. 1969. Logic of norms and logic of normative propositions. *Logique et analyse* 12: 242–268.

Alchourrón, C.E., and E. Bulygin. 1971. *Normative systems*. Vienna: Springer.

Belnap, N., M. Perloff, M. Xu, and P. Bartha. 2001. *Facing the future: Agents and choices in our indeterminist world*. Oxford: Oxford University Press.

Dung, P.M. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming, and *n*–person games. *Artificial Intelligence* 77: 321–357.

Governatori, G., and A. Rotolo. 2010. Changing legal systems: Legal abrogations and annulments in defeasible logic. *Logic Journal of IGPL* 18: 157–194.

Jones, A.J., and M.J. Sergot. 1996. A formal characterisation of institutionalised power. *Journal of the IGPL* 4: 429–445.

Kelsen, H. 1967. *The pure theory of law*. Berkeley: University of California Press.

Pörn, I. 1977. *Action theory and social science: Some formal models*. Dordrecht: Reidel.

Prakken, H. 2001. Relating protocols for dynamic dispute with logics for defeasible argumentation. *Synthese* 127: 187–219.

Prakken, H. 2010. An abstract framework for argumentation with structured arguments. *Argument and Computation* 1: 93–124.

Prakken, H., and G. Sartor. 1996. Rules about rules: Assessing conflicting arguments in legal reasoning. *Artificial Intelligence and Law* 4: 331–368.

Prakken, H., and G. Sartor. 2013. Formalising arguments about norms. In *Proceeding of JURIX 2013: The twenty-sixth annual conference on legal knowledge and information systems*, 121–30. Amsterdam: IOS.

Sartor, G. 2008. Legal validity: An inferential analysis. *Ratio Juris* 21: 212–247.

Sartor, G. 2012. Intentional compliance with normative systems. In *The goals of cognition. Essays in honor of Cristiano Castelfranchi*, ed. F. Paglieri, L. Tummolini, R. Falcone, and M. Miceli, 627–656. London: College Publications.

Searle, J.R. 1995. *The construction of social reality*. New York: Free.

Sergot, M.J. 2001. A computational theory of normative positions. *ACM Transactions on Computational Logic* 2: 581–662.

van der Torre, L., and J. Hansen. 2008. In *Deontic logic in computer science: Course material. 20th European summer school in logic, language and information*, Hamburg.

# An Agent Based Model of *Camorra*: Comparing Punishment and Norm-Based Policies in Contrasting Illegal Activities

**Barbara Sonzogni, Federico Cecconi, Giulia Andrighetto, and Rosaria Conte**

**Abstract** In this chapter, we will discuss the need of Agent Based Modelling (ABM) to study the dynamics of a specific type of illegal system, i.e., Extortion Racket Systems, which appear to be highly prosperous and to behave as a dynamic system, spreading wide and fast in current Western societies.

This work arises from two traditions of study: one is related to the deviance issue and the spread of organized crime, the other on the socio-cognitive study of norms. The strength of the realized simulation model is based on two factors: the reference to a previous empirical grounded model that it's complemented with the use of cognitively rich agents.

More specifically, we have implemented a case study, resembling as much as possible the Camorra phenomenon in Campania, aimed to test the relative and combined effect of punishment and norms in contrasting the spreading of Extortion Racket Systems. Results show that to be effective policies based only on punishment should be very severe. Nevertheless, when punishment is combined with norms, their effect in reducing the number of racket affiliates is not only stronger but also more stable over time with respect to punishment alone. These results enlighten the limits of the anti-crime strategies based merely on the use of punishment, and show the advantages of a multi-faceted policy that incorporates traditional, i.e., economic, and non-traditional, i.e., normative, factors.

B. Sonzogni (✉)
Department of Communication and Social Research, Sapienza University of Rome, Rome, Italy

Laboratory of Agent Based Social Simulation (ISTC – CNR), Rome, Italy
e-mail: barbara.sonzogni@uniroma1.it

F. Cecconi • R. Conte
Laboratory of Agent Based Social Simulation (ISTC – CNR), Rome, Italy
e-mail: federico.cecconi@istc.cnr.it; rosaria.conte@istc.cnr.it

G. Andrighetto
Laboratory of Agent Based Social Simulation (ISTC – CNR), Rome, Italy

European University Institute, Florence, Italy
e-mail: giulia.andrighetto@istc.cnr.it

A. Herzig, E. Lorini (eds.), *The Cognitive Foundations of Group Attitudes and Social Interaction*, Studies in the Philosophy of Sociality 5,
DOI 10.1007/978-3-319-21732-1_10

**Keywords** Extortion racket systems • Agent-based simulation • Rich cognitive agents • Social norms • Punishment

## 1 Introduction

The diffusion of illegality is one of the hardest problems of complex human societies. Illegality includes not only organized crime (extortion racketeering, illegal trafficking, and terrorism), but any violation of legal norms (such as financial frauds, corruption, cybercrime, tax evasion, private violence, etc.).

In the last decades, the diffusion of one specific form of organized crime has substantially increased: Extortion Racket Systems (from now on, ERSs) spread so fast and far, that we can now legitimately speak of ERSs as evolving systems, not only because they are highly prosperous, with revenues comparable to the Gross Domestic Product (GDP) of small countries, but also highly dynamic systems always in search for new markets to invest in and affluent societies to exploit (Forgione 2009; Varese 2011; La Spina 2008).

ERSs differ from sporadic extortion in that they aim to extract regular cash payments from a pool of economic agents and continue to do so over time. The victims are forced to pay by the threat of violence or other harmful retaliation and only if they pay will they suffer no harm. If they refuse to pay harm usually follows by incremental steps. What is extorted is normally a sum of money but it is also possible that the illegal organization steals goods or imposes given partners as unique suppliers to purchase from or as employees to recruit etc. A distinctive trait of ERSs is the monopolistic production and the forced supply of "protection" in exchange for money or other economically relevant utility (Schelling 1960; Gambetta 1993).

The fast growth of ERSs poses a challenging research question to the social scientists: why do ERSs spread so wide? What is the secret of their success?

To answer these questions, we need a new way of looking at ERSs as systems bringing about a sort of "social order" (Romano 1875; Olson 1993) based on a credible dominance hierarchy maintained through a consistent use of punishment, "protection" provision, and the delivery of social services – such as legal assistance to affiliates in jail and/or pensions to their relatives (Gambetta 1993). As geo-criminal maps (Forgione 2009) show a different distribution of organized crime in rich Western countries, cultural, economic, social and institutional factors must also play an important role.

To investigate the dynamics of ERSs, innovative instruments should be developed. First, we need instruments allowing hypotheses to be clearly formulated and tested experimentally. As this is difficult to be done in the real world, it may be done *in silico* by means, for example, of simulation methods, models and techniques.[1]

[1] The main objective of the European ICT FP7 Project GLODERS (2012–2015) (http://www.gloders.eu/) is to develop theory-driven set of computational tools to study, monitor, and possibly predict the dynamics of ERSs.

This work focuses on one specific ERS system: *Camorra*. Camorra represents a particularly efficient and modernised variant of *Mafia* that has established its control on *Campania*, the region around *Naples* (Di Gennaro and La Spina 2010). *Campania* is characterised by a constantly increasing demographic trend by no means matched by the productive capacity of the regional economic system or by the job opportunities available to the population. As a result of the combined effect of demographic trends and structural inadequacy of the economic system, organized crime started to be regularly and growingly fed by new labour force, leading to the replication of criminal families or *clans (Clan dei Casalesi, Alleanza di Secondigliano, Scissionisti di Secondigliano, Clan di Lauro, Lo Russo, Licciardi, Russo dei Quartieri Spagnoli, etc.),* which invaded the whole region in a relatively short period of time in the last quarter of the century.

In previous work (Sonzogni et al. 2011; Conte et al. 2010), a simulation model aimed to investigate the effect of legal punishment in limiting the spreading of ERSs has been developed. In line with the model of crime developed by Becker (1968), results showed that when punishment is not severe enough, racketeering increases as a linear function of the extortion level. Moreover, even if punishment works as a deterrent against extortion, the number of racket affiliates does not decrease linearly for increasing levels of punishment.

In this paper, we present an extension of Sonzogni et al. (2011) and Conte et al. (2010), aimed at testing the power of norms in limiting the spreading of ERSs. Norms provide a key mechanism for modifying people's conducts, and we are interested in testing their effect in reducing the expansion of ERSs and in comparing it with that which would be obtained by using policies based on mere punishment.

## 2 The Value Added of the Present Model

This work arises from two traditions of study, in which the research group has worked. One is related to the deviance issue (Sonzogni 2010a, b) and the spread of organized crime (Sonzogni et al. 2011; Conte et al. 2010), the other on the socio-cognitive study of norms (Conte et al. 2014; Andrighetto et al. 2007, 2010).

The strength of this model is based on two factors. We realized the present simulation model referring to a previous empirical grounded model presented in La Spina (2008) and Di Gennaro and La Spina (2010) and we complemented it using cognitively rich agents.

Concerning the first aspect, the model is inspired by real-world data with respect to the specification of the type and the mode of operation of the Camorra group, and to specifications of the economic-demographic context.

In order to calibrate the model, we have collected and analyzed data drawn from specific official sources (ISTAT, Eurispes, Department of Treasury, Bank of Italy) (Istat 2008, 2009, 2010; Alongi 1977) related to job and salary, regional and national accounts (national and regional GDP), taxation, rates of non-regular employment. The model mimics *in silico* a stylized framework of Campania concerning population dimension, rates of inactivity, employment, unemployment, and irregular job in the tertiary sector. The statistical data (considering the economic, demographic and employment variables) represent the structural part of the model. The tertiary sector is here represented because micro-firms and self-employed are likely victims of predation, more than the medium and large firms.

Concerning the second aspect, agents are endowed with a rich cognitive architecture, EMIL-A (Conte et al. 2014; Andrighetto et al. 2007, 2010, 2013), allowing them to recognize norms, detect their salience and dynamically update it, and finally to decide whether or not to comply with those norms. In our perspective, norms are prescribed conducts largely spread in a society (Conte et al. 2014). For a normative behavior to take place, the agents must first recognize the existence of the norm and to form the corresponding mental representations (namely, sets of beliefs and goals concerning the norm), then to reason and take decisions upon them. To take place this two-way process requires cognitively complex agents, like EMIL-A agents.

## 3 Model Description

The specific research question we aim to address is the following: which are the social and individual dynamics that drive individuals to become extorters, or vice versa to quit extortion-based activities?

We have implemented a case study, resembling as much as possible the Camorra phenomenon in Campania, aimed to test the relative and combined effect of punishment and norms in contrasting the spreading of ERSs.

In the simulation model, agents are assigned with one of the following *roles,* representing different types of possible economic activities: self-employed, employee, unemployed and racket affiliate.

Each agent has a specific role and at each step of the simulation automatically gets a *payoff* out of its honest (i.e., self-employed, employee) or illegal activity (i.e., racket affiliate). This payoff varies during the simulation as an effect of the economic dynamics. Agents can decide whether or not to quit working (thus becoming unemployed), to start a dishonest activity (thus becoming a racket affiliate) and to change their employment roles (moving from self-employed to employee, or vice versa).

The model explores the combined effects of three actions:

1. *Extort*. Racket affiliates extort self-employed agents (the owners of a micro-firm). Agents that are asked for extortion always pay the tribute.
2. *Punish*. With a certain probability and severity (see Sect. 4), an external agency – standing for the legal judiciary system – punishes agents undertaking illegal activities. At each simulation turn, punishment is inflicted (with a certain probability and severity) on racket affiliates thus reducing their payoffs.

3. *Change Role*. This decision is a function of *two* considerations: economic and normative ones. At the end of each simulation turn, each agent compares its own payoff with the payoff of all others agents. This comparison allows the agent to figure out which are the roles/activities that pay better. Moreover, each agent controls for the existence of the "no extorting norm" in its memory and checks for its salience. The higher the salience of the "no extorting norm", the higher its effect in refraining the agent from becoming a racket affiliate and undertaking an illegal extortive activity.

## *3.1 Normative Agent*

In the present model, agents are endowed with the normative architecture EMIL-A allowing agents to recognize which are the norms governing their interactions and to detect and dynamically modify the degree of salience of these norms. EMIL-A has three important components: the norm recognition module, the salience meter, which affects the norm adoption module and the formation of normative goals, and the norm compliance module, turning such goals into normative intended actions (see Fig. 1). More specifically, the norm recognition module allows EMIL-A agents to interpret a social input as a norm. Once the norm is recognized (in the present work the norm is the "no extorting norm"), agents generate a normative belief that will possibly activate a normative goal to comply with the abovementioned norm (for a detailed description of how the norm recognition module works, see (Conte et al. 2014)).

The salience meter *tells* the agent how salient a norm is. With salience, we refer to the perceived degree of prominence and strength of a norm within a given situation. Psychological evidence suggests that the more a norm is perceived as salient, the more likely it will be complied with (Cialdini et al. 1990). Norm salience is a parameter endogenously and dynamically updated at every simulation turn by each agent according to both the personal decisions taken by the agents and the normative and social information gathered by observing and communicating with others. For example, those behavioral or communicative acts that are interpreted as compliant with the "no extorting norm" (i.e., not paying extorters) or enforcing it (e.g., punishing racket affiliates) increase the salience of the norm. Conversely, violations (e.g., applying extortion), especially when unpunished, reduce the norm salience, by signaling that the group is losing interest in the norm (for a detailed description of how the norm salience is calculated and updated, see (Villatoro et al. 2011, 2013; Andrighetto et al. 2013)).

The norm compliance module allows the EMIL-A agents to compare the "no extorting" normative goal with other goals, e.g., the goal of maximizing its own payoffs, to choose which one to execute based on their respective salience values and convert it into an intention (i.e., an executable goal).

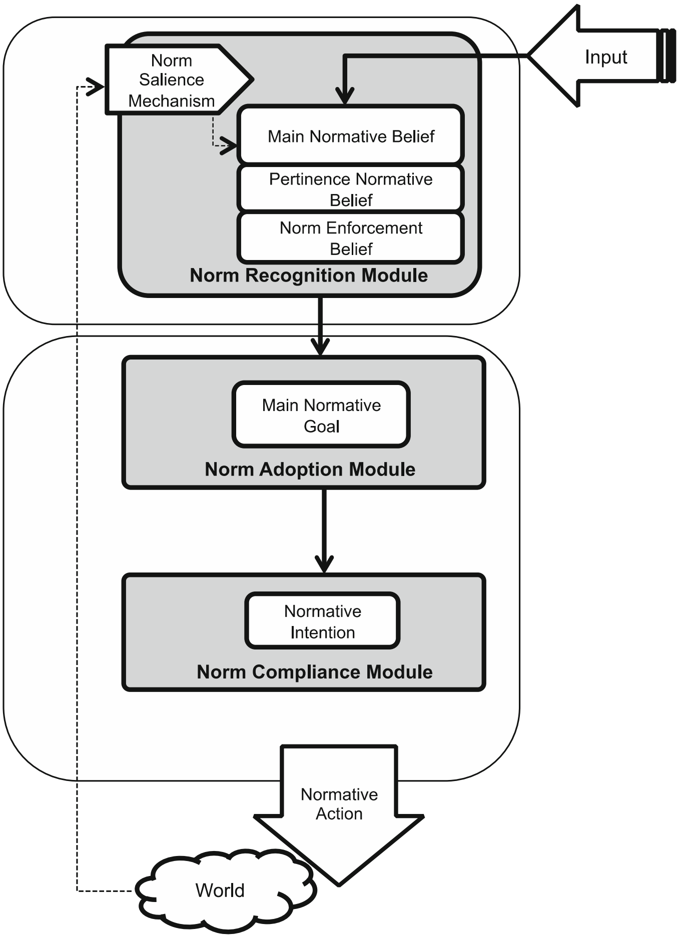


**Fig. 1** Main components and mental dynamics of EMIL-A: The architecture consists of different modules interacting with one another by means of input-output mechanisms. The norm recognition module plays a crucial role by informing both the norm adoption and the norm compliance modules. These two modules are responsible for the actions performed by the agent

## 4 Simulations: Description of the Experiments

The simulation model has been implemented using a NetLogo environment tool (Wilensky 1999). The simulation explores the interplay among the following variables: (Table 1)

The first set of simulation experiments is aimed to check for the effects of more or less severe punishment policies in limiting the spreading of extortive activities. The second set of experiments is aimed to test for the effect of combining punishment-based with norm-based policies. Finally, the last set of simulations is aimed to test

**Table 1** Independent and dependent variables analyzed in the simulation

| Independent variables | Dependent variables |
|---|---|
| Level of extortion (*LevExtortion*): payoff share of a self-employed that is extorted by a racket affiliate | Number of self-employed |
| Level of punishment (*LevPunish*): payoff share of a racket affiliate that is punished | Number of employees |
| Extortion probability (*ExtortionProb*): how often racket affiliates extort self-employed | Number of unemployed |
| Punishment probability (*PunishProb*): how often racket affilates are punished | Number of racket affiliates |
| Normative reasoning: flag that indicates if normative considerations are taken into account when deciding which role to adopt | Number of extorted agents |
| | Number of punished agents |
| | Payoff: average wage for all different status |
| | Norm salience |

the hypothesis that norm-based policies produce a more stable effect in controlling the expansion of extortive activities than punishment-based strategies alone.

# 5 Results and Discussion

## *5.1 Extortion Dynamics in Presence of Punishment and Norms*

The first round of experiments is aimed to test the effect of more or less severe punishment policies in limiting the spreading of extortive activities. In this first experiment, the probability of a racket affiliate to be punished is equal to 50 % (punishment probability = 0.5); the probability that a racket affiliate extorts a self-employed agent is equal to 50 % (extortion probability = 0.5); the extortion consists in a 50 % lowering of the payoff of the victim (level of extortion = 0.5); and the normative reasoning is not active, meaning that norms do not affect agents' decisions that are driven only by utility-based considerations (normative reasoning = 0).

In Fig. 2, the line series represent the different levels of punishment (i.e., punishment at different degrees of severity 0 % 20 %; 50 %; 80 %; 100 %) inflicted on the racket affiliates by the external agency; the vertical axis shows the average amount of racket affiliates; the horizontal axis represents the simulation steps. Results are based on 30 simulation replications.

Figure 2 shows that in two (0 %, 20 %) out of five punishment conditions, extortive activities increase over time. If punishment severity is at 50 %, the number of people involved in racketeering remains stable.

Punishment works as a deterrent against extortion only when it is highly severe (level of punishment = 80 % and 100 %).

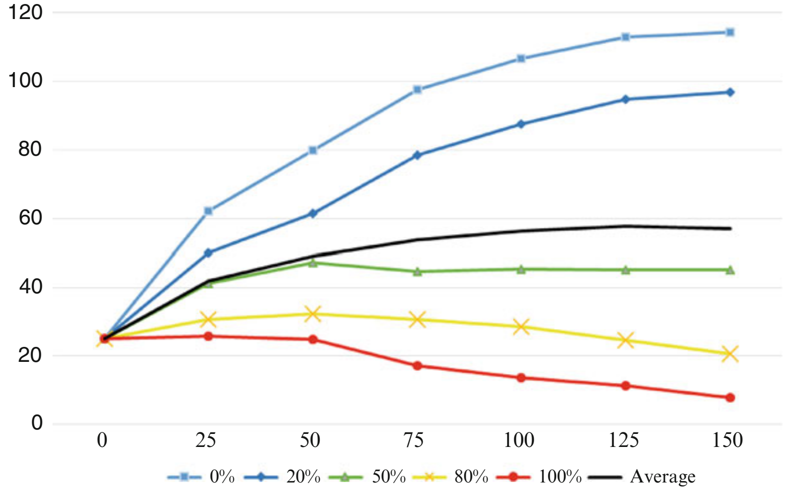


**Fig. 2** Effect of more or less severe punishment policies on the number of racket affiliates (no normative reasoning). On the x axis the number of simulation steps is shown; on the y axis the average amount of racket affiliates

How the decision of the agent to change its role and start undertaking an illegal extortive activity is affected by the presence of norms?

We run a second set of simulations in which the normative reasoning of the agents is active (normative reasoning = 1). As in the previous experiment, the probability of a racket affiliate to be punished is equal to 50 % (punishment probability = 0.5); the probability that a racket affiliate extorts a self-employed is equal to 50 % (extortion probability = 0.5); and the extortion consists in a 50 % lowering of the payoff of the victim (level of extortion = 0.5).

In Fig. 3, the line series represent the different levels of punishment (i.e., punishment at different degrees of severity 0 %, 20 %; 50 %; 80 %; 100 %) inflicted on the racket affiliates by an external agency; the vertical axis displays the average amount of racket affiliates; the horizontal axis shows the simulation steps. Results came out from simulation runs replied 30 times.

Results show that combining norms with punishment is highly effective in reducing the amount of racket affiliates. We observe that in the long run even with no (0 %) or low punishment severities (20 %) the introduction of norms allows to reduce the number of racket affiliates with respect to the situation in which only punishment is used (compare Figs. 2 and 3). Agents having the “no extorting norm” in their minds have an additional reason for refraining from extortion, other than only avoiding punishment.

Since the aim of this work is to explore the combined effect of punishment and norms, the normative reasoning has been activated in each agent in all the experimental conditions, even in the presence of low punishment. We have then run a second set of experiments (see Fig. 4) in which the norm salience is highly

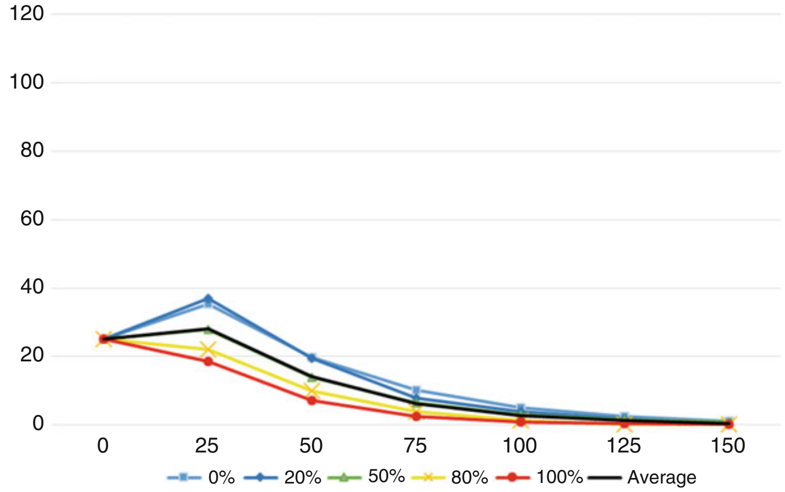


**Fig. 3** Effect of more or less severe punishment policies on the number of racket affiliates (with normative reasoning). On the x axis the number of simulation steps is shown; on the y axis the average amount of racket affiliates

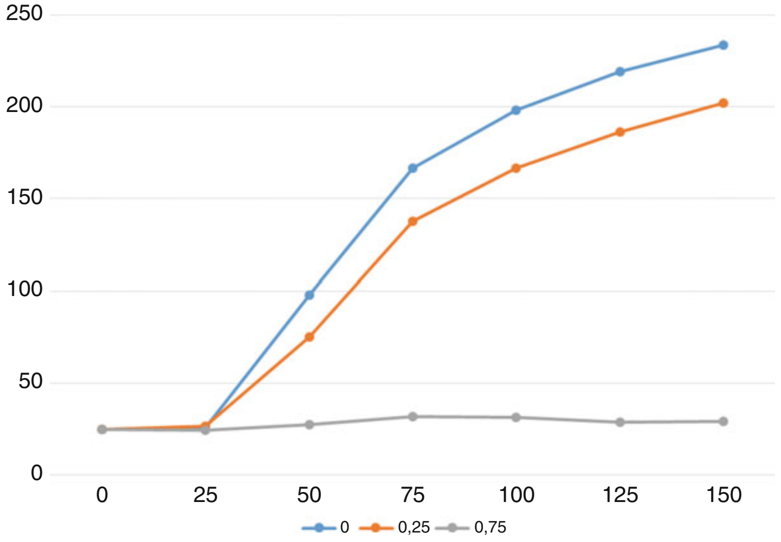


**Fig. 4** Effect of more or less severe punishment policies on the number of racket affiliates (with normative reasoning and with norm salience updating strongly impacted by norm violations). On the x axis the number of simulation steps is shown; on the y axis the average amount of racket affiliates

sensitive to norm violations and is strongly affected by the observation of illegal acts. It is sufficient that the agent observes few illegal acts that the salience of the

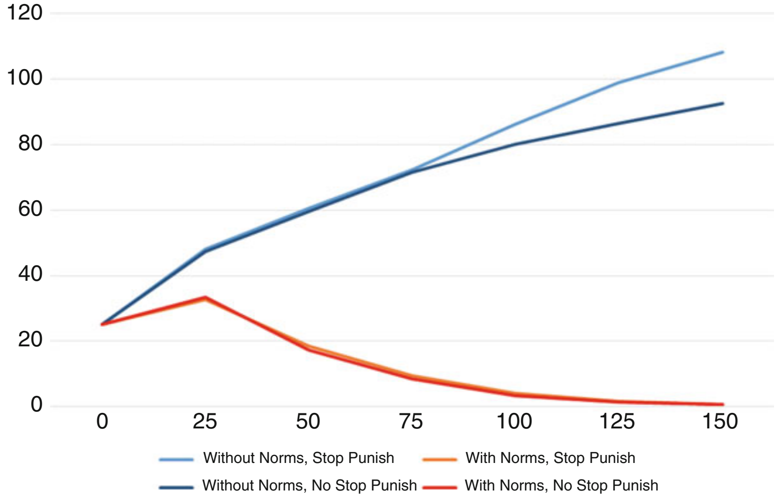


**Fig. 5** Effect of interrupting punishment on the number of racket affiliates. The results of the two conditions, with and without normative reasoning, are shown

"no extorting norm" quickly decreases. As shown in Fig. 5, with these new initial parameters for the norm salience activation and updating, if punishment is not severe enough, the normative reasoning is not even activated.

## 5.2 *What Does It Happen When Punishment Is Interrupted?*

In the third experiment, after step 25, punishment has been interrupted in the two conditions, with and without normative reasoning.

Figure 4 shows that by suddenly interrupting punishment the number of racket affiliates increases more in the condition in which the normative reasoning is not active with respect to the situation in which the normative reasoning is active. The difference in the number of racket affiliates observed by comparing the two conditions "without norms, stop punish" and "without norms, do not stop punish" is higher than the one obtained by comparing the two conditions "with norms, stop punish" and "with norms, do not stop punish".

Agents that have generated the "no extorting norm" in their minds (from simulation step 1–25) are less prompt to switch from honest to dishonest activities even when deterrent penalties are interrupted than agents with no norms. Combining norm-based and punishment-based policies has a positive effect in reducing the number of extortive activities and in making this result stable over time.

## 6 Conclusions

In this paper, we have discussed the necessity of Agent Based Modelling (ABM) to study the dynamics of a specific type of illegal systems, i.e., Extortion Racket Systems, which appear to be highly prosperous and dynamic systems, spreading wide and fast in current Western societies.

The more credible and stable the dominance system they establish, the more prosperous they get and the likelier they are to move to new territories in search of new investments. As a consequence, ERS tend to replicate, new variants appear and compete on the same territory or move to the conquest of new ones.

An ABM-based study of Camorra in Campania is presented. The simulations allowed us to observe the relative and combined effect of punishment-based and norm-based policies in fighting against the spread of ERSs. Results show that to be effective policies based only on punishment should be very severe. Nevertheless, when punishment is combined with norms, their effect in reducing the number of racket affiliates is not only stronger but also more stable over time with respect to punishment alone. These results enlighten the limits of the anti-crime strategies proposed by Backer (1968), based merely on the use of punishment, and show the advantages of a multi-faceted policy that incorporates traditional, i.e., economic, and non-traditional, i.e., normative, factors.

We aim to run new simulation experiments to explore further the advantages of combining top-down with bottom-up interventions for fighting against the spreading of ERSs.

**Acknowledgments** This work was supported by the FP7 European ICT Project GLODERS (315874).

## References

Alongi, G. 1977. *La maffia*. Palermo: Sellerio.

Andrighetto, G., M. Campennì, R. Conte, and M. Paolucci. 2007. On the immergence of norms: A normative agent architecture. In *Proceedings of AAAI symposium, social and organizational aspects of intelligence*, Washington, DC.

Andrighetto, G., D. Villatoro, and R. Conte. 2010. Norm internalization in artificial societies. *AI Communications* 23: 325–339.

Andrighetto, G., J. Brandts, R. Conte, J. Sabater, H. Solaz, and D. Villatoro. 2013. Punish and voice: Punishment enhances cooperation when combined with norm-signalling. *PLoS One* 8(6): 1–8.

Becker, G. 1968. Crime and punishment: An economic approach. *Journal of Political Economy* 76(2): 169–217.

Cialdini, R.B., R.R. Reno, and C.A. Kallgren. 1990. Focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology* 58(6): 1015–1026.

Conte, R., F. Cecconi, and B. Sonzogni. 2010. Dynamics of illegality. The case of Mafia systems. In *ECCS'10 Lisbon European conference on complex systems*, Lisbon.

Conte, R., G. Andrighetto, and M. Campennì. 2014. *Minding norms. Mechanisms and dynamics of social order in agent societies*, Oxford series on cognitive models and architectures. Oxford: Oxford University Press.

Di Gennaro, G., and A. La Spina (eds.). 2010. *I costi dell'illegalità. Camorra ed estorsioni in Campania*. Bologna: Il Mulino.

Forgione, F. 2009. *Mafia export. Come 'Ndrangheta, Cosa Nostra e Camorra hanno colonizzato il mondo*. Milano: Baldini Castoldi Dalai Editore.

Gambetta, D. 1993. *The Sicilian Mafia: The business of private protection*. Cambridge: Harvard University Press.

Istat. 2008. Rilevazione sulle forze di lavoro, Media 2007.

Istat. 2009. La misura sommersa secondo le statistiche ufficiali Anni 2000–2008.

Istat. 2010. Dossier. L'economia sommersa: stime nazionali e regionali.

La Spina, A. (ed.). 2008. *I costi dell'illegalità. Mafia ed estorsioni in Sicilia*. Bologna: Il Mulino.

Olson, M. 1993. Dictatorship, democracy, and development. *The American Political Science Review* 87(3): 567–576.

Romano, S. 1875–1947. *Frammenti di un dizionario giuridico*. Milano: Giuffrè.

Schelling, T. 1960. *The strategy of conflict*. Cambridge: Harvard University Press.

Sonzogni, B. 2010a. Percorsi penali. Dalla concettualizzazione sociologica alla penologia di contenimento della recidiva. In *Recidività e reinserimento. L'affidamento in prova al servizio sociale nel Lazio*, ed. M. Bonolis, 205–228. Acireale-Roma: Bonanno.

Sonzogni, B. 2010b. Reati e progetti su misura: possibilità e limiti degli interventi di servizio sociale. In *Recidività e reinserimento. L'affidamento in prova al servizio sociale nel Lazio*, ed. M. Bonolis, 243–264. Acireale-Roma: Bonanno.

Sonzogni, B., F. Cecconi, and R. Conte. 2011. On the interplay between extortion and punishment. An agent based model of Camorra. In *2011 Computational Social Science Society of America annual conference (CSSSA 2011)*, Santa Fe, New Mexico.

Varese, F. 2011. *Mafias on the move: How organized crime conquers new territories*. Princeton: Princeton University Press.

Villatoro, D., G. Andrighetto, R. Conte, and J. Sabater-Mir. 2011. Dynamic sanctioning for robust and cost-efficient norm compliance. In *Proceedings of the twenty-second International Joint Conference on Artificial Intelligence (IJCAI 2011)*, pp. 414–419.

Villatoro, D., G. Andrighetto, J. Brandts, L. Nardin, J. Sabater-Mir, and R. Conte. 2013. Norm signalling effects of group punishment. *Social Science Computer Review* 32: 334–353.

Wilensky, U. 1999. NetLogo. http://ccl.northwestern.edu/netlogo/