Communicating cognitive agents: modeling and formalization

Dominique Longin

Habilitation à diriger les recherches

Communicating cognitive agents: modeling and formalization

Dominique Longin

Dominique.Longin@irit.fr
www.irit.fr/∼Dominique.Longin/

Sous la direction de Andreas Herzig

JURY

Reporters:

Marco Colombetti, Pr. in Computer Science, Università degli Studi di Milano
Shahid Rahman, Pr. in logic and epistemology, Université de Lille
Laurent Vercouter, Pr. in Computer Science, INSA Rouen

Examinators

Nicholas Asher, DR in Computer Science, CNRS
Marie-Pierre Gleize, Pr. in Computer Science, Université Toulouse 3
Andreas Herzig, DR in Computer Science, CNRS (supervisor)

Université de Toulouse
CNRS, Université Toulouse 3 Paul Sabatier
Toulouse, France

March 27, 2015

Abstract

This work concerns cognitive agents, i.e. agents with mental states such as belief, goal, intention, internalized norms, etc. These mental states can be about facts of the world (including laws of action), about their own mental states, or about the mental states of other agents. Cognitive agents are able to reason about all their mental states.

In what follows, we first give an overview of our contribution (Chapter 2), and then more details on the three most important parts of this contribution (Chapter 3). The first part concerns individual cognitive agents; the second part concerns social concepts (that is to say the groups of agents seen in a global way); and finally, the third part concerns the modeling of emotions. In the next chapter (Chapter 4) we list the publications that are related to each part of the previous chapter.

Contents

1 Introduction **3**
- 1.1 What is a cognitive agent? . . . 3
- 1.2 Are mental states and action the right concepts for cognitive agents? . . . 5
  - 1.2.1 Mental states . . . 5
  - 1.2.2 About doxastic and epistemic mental states . . . 8
  - 1.2.3 About motivational mental states . . . 10
  - 1.2.4 About normative mental states . . . 16
  - 1.2.5 About action . . . 17
- 1.3 Short history of cognitive systems . . . 21
- 1.4 Conclusions . . . 24

2 Overview **26**
- 2.1 Research themes . . . 26
- 2.2 History and not detailed contributions . . . 28
  - 2.2.1 PhD thesis (1995–1999) . . . 28
  - 2.2.2 Multidisciplinary contribution around language (1999–2008) . . . 32
  - 2.2.3 After PhD Thesis . . . 34
- 2.3 Conclusions . . . 36

3 Summary of selected articles **37**
- 3.1 Mental attitudes and their dynamics . . . 37
  - 3.1.1 Non reductionist view of intention . . . 37
  - 3.1.2 Reductionist view of intention . . . 43
  - 3.1.3 Sensing actions, belief change and misperception . . . 47
- 3.2 Social concepts . . . 53
  - 3.2.1 Groundedness as the expression of beliefs . . . 55

3.2.2 The concept of acceptance . . . 63
3.3 Emotion and trust . . . 72
3.3.1 Emotion . . . 72
3.3.2 Trust . . . 81

4 Selection of published articles **85**
4.1 Speech acts and the dynamics of mental states . . . 85
4.1.1 A logic of Intention with Cooperation Principles and with Assertive Speech Acts as Communication Primitives (AAMAS'02) . . . 86
4.1.2 C&L Intention Revisited (KR'04) . . . 94
4.1.3 Sensing and revision in a modal logic of belief and action (ECAI'02) . . . 103
4.2 Social concepts . . . 108
4.2.1 A Logical Framework for Grounding-based Dialogue Analysis (ENTCS 2006) . . . 109
4.2.2 A new semantics for the FIPA Agent Communication Language based on Social Attitudes (ECAI 2006) . . . 130
4.2.3 The logic of acceptance: grounding institutions on agents' attitudes (JLC 2009) . . . 135
4.2.4 A logical account of institutions: from acceptances to norms via legislators (KR 2008) . . . 183
4.3 Emotion and trust . . . 194
4.3.1 A logical formalization of the OCC theory of emotions (Synthese 2009) . . . 195
4.3.2 The face of emotions: a logical formalization of expressive speech acts (AAMAS 2011) . . . 244
4.3.3 A Logical Framework for Trust-Related Emotions (EASST 2009) . . . 252

5 New research perspectives **267**
5.1 Introduction . . . 267
5.2 Génèse . . . 268
5.3 Cadre sociétal et enjeux . . . 269
5.4 Cadre scientifique . . . 270
5.5 Programme de recherche et contributions à l'état de l'art . . . 274
5.6 Conclusions . . . 278

Chapter 1

Introduction

1.1 What is a cognitive agent?

The general theme of the work presented in what follows is that of intelligent agents. Many definitions can be given on what an intelligent agent is, and this requires first of all to define what an agent is. It is obviously true that one can find a very large set of definitions in the literature, these definitions being often very close, sometimes contradictory. In the following, an agent is defined as an entity having **properties** such as:

- autonomy (the ability both to act without human intervention and to control one's actions and internal states);
- reactivity (the ability to interact with other agents *via* a communication language);
- proactiveness (the ability to adopt goal-directed behavior by taking the initiative to do something);
- etc.

Sometimes the *properties* themselves can also be more specific to humans. For example:

- rationality (in a very general sense, a rational agent does not act in a contradictory way: he does not believe both something and its opposite, he acts in accordance with his goals, etc.);

- sincerity (an agent is sincere when he doesn't want to cheat on someone);
- etc.

Following Wooldridge, agents "are able of *deciding for themselves* what to do in any given situation" [160].

It is interesting to note that these properties depend on the universe in which the agents evolve. [1] For example, is it appropriate to assume that an agent must be sincere when intended to play poker? Is it more appropriate to assume that he is not sincere when intended to communicate a weather forecast for the holidays? Probably the answer is "no" in both cases.

In the following, cognitive agents are agents for which the above properties are described using concepts generally associated with humans:

- mental attitudes (belief or knowledge, goal or desire, intention, etc.);
- social attitudes (commitment, collective belief or intention, acceptance, etc.);
- norms, institutions, time, physical or linguistic actions, etc.

When the above properties are used by designers to implement a particular system, that system is named "cognitive agent" (when it describes only one agent) or "system of cognitive agents" (when several cognitive agents are described) with the aim of clarifying that this system is built from concepts specific to man. Thus, the behavior of such systems should be predictable according to the mental attitudes assigned to them. Therefore, the problem is then to choose the "right" mental attitudes for a given system.

In addition, the cognitive agent must be able to reason. Reasoning is in fact the core of a cognitive agent's intelligence. Mental attitudes are just the way properties are represented in this agent. But properties such as *modus ponens* or *rationality* distinguish cognitive agents from reactive agents.

The question that remains is: why do we need cognitive agents? Maybe the correct question should be "why do we need reactive agents?" and the answer should be: as long as the expected behavior of an agent can be easily described by the system designer in any situation, reactive gents are both

[1] Wooldridge says that the agents should be "*embodied* in some environment" that is to say that 'they should "inhabit and act upon some environment in the same way we inhabit and act upon ours".

a simple and very efficient solution. So reaction of reactive agents is just a built-in function of *stimuli* (something like a table in a database system where a particular value in the first column gives the corresponding value -or, reaction- in the second column).

But as soon as the task becomes complex, it is very difficult for the designer to calculate every reaction of the agent in any situation. It is therefore useful to define more elaborate concepts (such as emotion, delegation, trust, etc. for example, but also sincerity, rationality, etc.), facilitating the description of the system and its formalization. In addition, it is often useful to explain the reaction of an agent in a particular situation, which is impossible when this reaction is directly calculated in the agent: in reactive agents, reasoning leading from a stimulus to agent reaction is not available.

1.2 Are mental states and action the right concepts for cognitive agents?

To sum up the preceding section, in the following cognitive agents are particular agents having mental states, able to reason about these mental states and to act in accordance with the conclusion of their reasoning. Thus, mental states and action are at the core of these works. But what are mental states? What is action? And why are these concepts both interesting and appropriate for cognitive agent modeling? Thus, we propose in the rest of this chapter to present an overview of these concepts such as we intend to use them.

1.2.1 Mental states

Mental states of a given agent (such as: belief, intention, desire, thought, etc.), in contrast with physical states, are states that are only accessible to this agent.[2]

We can distinguish, among others, three main types of mental states: epistemic mental states such as belief and knowledge; motivational mental states such as desire, goal, intention, etc.; normative mental states such as internalized norms (moral values, obligations, duties, etc.). Epistemic states

[2]There is a great debate for knowing how physical states and mental states are related. See [118] for instance for an overview of this question.

allow agents to represent their (subjective perception of their) environment, to describe what is true or false from their point of view. Motivational states are used for describing the world such as agents would like it to be. Finally, normative states are used for describing the world such as it should be from the point of view of agents according to some standards. Agent models based on belief, desire and intention are commonly called "BDI architectures" in the literature.

In the following we only consider mental states having *Intentionality* that is, related to or about things in the real world. We follow here the point of view of Searle who says that "Intentionality is that property of MANY mental states and events by which they are directed at or about or of objects and states of affaires in the real world." [137, p. 1] (emphasis added).[3] The reason is that such non Intentional states are generally not considered in AI and we do not deal with them in the following. In other words, mental states can be represented by mental attitudes (belief, desire, intention, etc.) applying to an Intentional propositional content. Formally, we will write for instance: $Bel_i\ skyIsBlue$ (that is read: agent i believes that the sky is blue) or $Des_j\ getsMoreMoney$ (agent j desires to get more money). Note that this does not presuppose that the object of an Intentional mental state exists in the real world. For instance, I can believe that the Sara's car is blue, even if in the real world Sara has no car at all. Finally, note that intention is just a particular Intentional mental state (that is why, as Searle, we distinguish intentionality from Intentionality by the major caps).

As noted by Searle [137], every Intentional concept has a direction of fit that gives the *conditions of satisfaction* of these concepts, that is, the necessary conditions for having the object of an Intentional concept true. In fact, the direction of fit of mental states follows from their Intentionality and describes how these mental states are related to their object. For instance, epistemic mental states have a mind-to-world direction of fit whereas motivational states have a world-to-mind direction of fit. It means that the object of an epistemic state is a true proposition when the mind fits to the world (the agent believes something that is true in the world). When this is the case we say that this mental attitude is *satisfied*. In the same way, the object of a motivational state is true when the world changes so that the object

[3]It is not obvious for every philosopher that there exist non Intentional mental states but Searle does not agree with them and he cites some "forms of nervousness, elation, and undirected anxiety" [137, p. 1] that are not, from his point of view, Intentional [137, p. 6–7].

describes the world as it is now and we say that this motivational state is satisfied when the world has changed in a way described by the object of this state.

Finally, mental states can be conscious or unconscious. We can both believe that the sky is blue and use this belief for choosing our clothes without be conscious of that belief (we are in summer and the sky has been blue every day for several weeks).

It is interesting here to compare this formalization of mental states with works in cognitive psychology. Some theories describe cognition by declarative memory (DM) on the one hand, and by procedural memory (PM) on the other hand (see [9, 10] for instance). DM is the set of facts (pieces of information called "chunk") that we are aware to know whereas DM is a set of know-how (a set of rules allowing reasoning about chunks). "Aware" does not mean here that DM contains only facts we are conscious of: we can have forgotten the fact that p is true even if p is in our DM, but if it is needed we can (attempt to) retrieve p from DM. PM is the set of all low-level mental events, a little like the low-level instructions set of a microprocessor. What is interesting in this conception of our memory is that DM does not make any distinction between a chunk related to a goal and a chunk related to belief: as a microprocessor needs registers for functioning, PM needs what Anderson calls *buffers*. There exists several buffers (such as goal buffer or visual buffer for instance) and a chunk is assimilated to a goal when it is in the goal buffer; when a chunk is in the visual buffer, it is considered to be the result of a visual observation. Thus, mental states exist only in central cognition[4] when our attention focuses on a particular task. In AI, we generally do not make this distinction and we only handle mental states all the time.

The majority of philosophers such as Searle [138] or Dennett [43] for instance think that mental states follow in some sense from brain activity.[5] For Dennett, every mind activity can be reproduced by a machine (thus, the brain is not a necessary thing for having thoughts and consciousness: we can simulate it by a digital machine). This point of view has been adopted by what is called strong IA (see McCarty's works for instance, where every

[4]The part of the cognition where chunks are used for reasoning.

[5]The goal is to give an unified view of the brain and of the mind. This is an important question for philosophers because they need to explain the relations between the brain/body and the mind. (There also exists a dualistic point of view where brain and mind are considered as different things; see plato's and Descarte's works for instance, and [127] for an overview of the dualism in philosophy of mind.)

thing can be described by the help of the mental states model). But Searle criticizes this view [138, 134] and argues that computers cannot have mental states in the same sense as humans: for him, computers are just pure symbolic machines whereas semantical machines would be needed to capture human cognition (thus, unlike Dennett, he think that the brain is necessary to thought). In other words, event if a computer can deal with mental states in some syntactical way, we should not say that computers have mental states (see the Searle's famous Chinese room).

There is always a strong debate in philosophy for or against Searle's and Dennett's points of view. But from the point of view of computer sciences, maybe it is not important if a machine does not have really mental states as soon as this machine behaves *as it had mental states*, as soon as it complies with the human social rules by performing its task. Thus, the remaining question is: does there exist a behavior that a computer could have only if it was a semantical machine? This is a very hard epistemological question that is clearly out of the scope of the present work.

Mental states have been identified very soon in AI as a common way for modeling a cognitive agent. (McCarthy said that even simple machines such as a thermometer has beliefs.) In fact, it is very useful for computer scientists to model a system with mental states: they can thus explain why an agent has such belief or such desire, and what impact can these mental states have on other mental states and on agent's actions (causality). Thus, the adopted point of view in AI is often close to Functionalim: mental states are described by the way of their function in a given system.[6]

1.2.2 About doxastic and epistemic mental states

We present here some view on belief states. They are particularly important in our works. By "important" we mean that every cognitive agent should have belief states for representing its subjective view of the real world. It does not mean, as it has been pointed out by Searle [137, pp. 29–36], that belief (and it is the same with desire) is something like an atomic mental state that would be needed in every complex mental state. Of course, a lot of mental states are built with the help of belief but not all.

[6]Maybe the reason is that Bratman's works have had a lot influence on AI models of intention and his answer to the question "What is intention?" is "broadly speaking within the functionalist tradition in the philosophy of mind and action." [22, p. 15].

What is the difference between belief and knowledge? Basically, following Kant, belief is traditionally formalized in modal logic as subjective knowledge [87]. It means that belief does not concern something, from the point of view of an agent, that is just *possibly true* in the real world, but something that *is true* for it in the real world (even if it is not really the case[7]). For instance, if agent i believes that Cannelle is a cat, it means that from the point of view of i Cannelle is a cat and i has no doubt about this fact. In particular, it does not envisage the possibility that Cannelle is not a cat even if in the real world Cannelle is a dog. Thus, from a subjective point of view, the object of the belief (here, the proposition *Cannelle is a cat*) is currently true in the real world.[8]

Thus, knowledge is often viewed as a true belief and defined from it. (Formally, we often have the following property: $Know_i\, p \rightarrow Bel_i\, p \wedge p$.) In other words, if an agent knows that something is true then this thing is necessarily true in the real world: I cannot know a fact whereas this fact is false in the real world. The main criticism of this view of knowledge is the mix-up between both subjective and objective component. From an intuitive point of view, we should not be able to know if what we believe is a piece of knowledge or a piece of belief. Thus, some philosophers have proposed that knowledge should be defined as a *justified* true belief rather than just a true belief. But a problem remains: what is a justified belief when an agent can have wrong belief about justifications of her/his knowledge? (See [65] for more details.)

Traditionally [48] the standard doxastic logic (that is, the logic of belief) is KD45 while the standard epistemic logic is S5. The both are normal modal logic and suppose that agents are rational (Axiom D meaning that the belief/knowlege of an agent cannot be contradictory) and are conscious of their belief/knowledge (Axioms 4 and 5 together with D entail that an agent believes/knows something if and only if it believes/knows that it believes/knows this thing). In real life, we are not always conscious of what we (do not) believe and both axioms 4 and 5 may be criticized. Nevertheless, such criticisms are justified only in cases where we want to stop some reasoning processes in the agent cognition. But it concerns only very specific situations. Of course, there exist some other logical frameworks that are

[7]Thus, if we note $Bel_i\, p$ the fact that agent i believes that p is true, $\neg p \wedge Bel_i\, p$ is consistent.

[8]Such belief is sometimes called *strong belief* in contradistinction to *weak belief* that is some kind of subjective probability or belief with degree of strength.

more or less different depending on the needs of authors.

1.2.3 About motivational mental states

Concerning motivational mental states, it is often hard to understand what exactly is captured because a lot of terms are used in literature (desires, goals, intentions, etc.): sometimes several terms are used for the same concept, and sometimes several concepts are named by the same term. In our terminology (see [2] for instance), a desire is a very primitive kind of motivational mental state: we can have contradictory desires and we can desire something that we believe impossible. Desires can also conflict with our moral values (see Section 1.2.4).

Goals are either chosen desires or chosen norms. They cannot be contradictory and we necessarily believe (or expect) that they can be achieved. In other words, they can be viewed as a rational part of chosen desires (or norms) that we would like to be true. Contrary to desires, goals are pro-active mental states because they are the first step towards action. Goal have traditionally been formalized in a logical framework of type KD (see [32, 122] for instance) and sometimes in a KD45 framework (see [129, pp. 83–84] for instance).

Finally, intention is certainly the most used concept and it is necessarily to give some details here.[9] When we deal with cognitive agents (but not only) we necessarily deal with intention because intention is, roughly speaking, the link between mental states and actions.

About (the three kinds of) intention. Certainly the first major work on intention is a book having this named written by Anscombe in 1957 [12, p. 1] where three kinds of intention are distinguished:

- *Future-directed intentions* are expressed in sentences such as in "I intend to swim in a club this year [but I do not have downloaded the registration form yet]". This is a kind of prospective intention that does not require to

[9]Note that, contrarily to Davidson for instance, some authors (following Anscombe [12]) do not view intention as a mental state but as a property of action itself. It is hard to understand what are the properties of intention that do not apply when intention is a mental state instead of something else, and the great majority of authors consider that intention is a mental state. We adopt this point of view without entering more in this debate here.

have begun to act for fulfilling it. If, as noted by Bratman, "we use the concept of intention to characterize both our actions and our minds" [22, p. 15], future-directed intention characterizes the mind rather than the action itself because it does not really speak about action itself, but it speaks about the mental state of the agent here and now. (Maybe the agent will never do this action.) This is the reason why this kind of intention is sometimes called "pure intention", that is, *pure* of any present execution.

• *With intention* is related to the fact that one acts with a certain current intention. For instance, I am going to Paris with the intention to see my brother. This kind of intention is strongly related to an explanation of the action itself: why am I going to Paris? Because I want to see my brother. This kind of intention seems to be like a present motivation during the execution of the action. Here (and contrarily to future directed intention) intention concerns the action itself (*versus* mental states).

• Finally, *intentional action* concerns the fact that we act intentionally. For instance, at this moment I am writing these words intentionally. In contrast of the pure intention, we could say that intentional action speak about a present directed intention, what I do here and now. Of course we can say sentences like "I did it intentionally" and we do not refer in this case to now but to the present time in the past, when I performed intentionally the action.

The main difficulty of philosophy of intention is to develop a theory unifying these three kinds of intention. It is not easy because some properties seem to be applicable to one kind of intention but not to the others. For instance, the fact that the intended action should be done in the future applies for future directed intention but not for the two other kinds of intention.

In a reductionist way, intention is generally viewed as a complex expression of beliefs (and sometimes goals). It was the point of view of the *first* Davidson for instance [39] which has noted that "Whenever someone does something for a reason (...) he can be characterized as (a) having some sort of pro attitude toward actions of a certain kind, and (b) believing (or knowing, perceiving, noticing, remembering) that his action is of that kind." He calls this complex of pro-attitude and belief the "primary reason" of the action. For instance, the fact that I intentionally turn the light on can be reduced to the facts that: a) I have the goal to turn on the light and b) I believe that some particular body movements correspond to such action of turning the

light on. By giving this reason, Davidson says that he gives the *intention with* which I turn the light on: "To know a primary reason why someone acted as he did is to know an intention with which the action was done."[10] [39]. The underlying hypothesis was: if these reasons are sufficiently detailed and specific, they characterize the action itself. Moreover, Davidson says that "a primary reason for an action is its cause". He illustrates that point with the help of the Melden's driver example. A man is driving his car but has no indicator anymore. Thus, in order to signal he raises his arm. But when he raises his arm it is possible he does not signal (maybe he greets somebody). Thus, if his action is to signal, then his primary reason must be his intention to signal and this is this intention (and not another one) that must causes his action to raise his arm. Thus, the caused action is necessarily intentional. In other words, in Davidson's mind, he has unified both together the *with intention* and *intentional action* concepts.

Example 1.1: Davidson's climber example [40, p. 79]

"A climber might want to rid himself of the weight and danger of holding another man on a rope, and he might know that by loosening his hold on the rope he could rid himself of the weight and danger. This belief and want might so unnerve him as to cause him to loosen his hold, and yet it might be the case that he never chose *to loosen his hold, nor did he do it intentionally."*

But just a simple causal link is not sufficient. Searle takes the example of raising one's arm [137, p. 85] and asks himself what could be the content of the corresponding prior intention. It cannot be that the arm goes up because it can go up without one intends that it go up (for instance, somebody else moves our arm up). In the climber's example (see Example 1.1) Searle remarks that the climber could intend to loosen his hold and that intention could "make him so nervous that he loosens his hold unintentionally" [137, p. 83]. Thus, the content of the prior intention for raising our arm cannot just be the fact that our intention causes the arm raising. It cannot be the action itself to raise the arm (because we can intend to raise our arm for a given reason but having this intention, we raise our arm for another reason —for instance, somebody ask who wants 10 millions dollars and I raise my

[10]More precisely, Davidson says that the primary reason entails the intention, and that it is not necessary to describe the entire primary reason for describing the entire intention.

arm to say: "I want 10 millions"). Thus, Searle proposes that intention must be linked to an action in a both together causal and self-referential manner. Thus content of one intention is that one raises the arm as a result of one's intention. (We intend to raise our arm in virtue of the fact that we have this intention.)[11]

In [41], Davidson changes his mind and speaks about intention as a full requirement for explaining action. Moreover, he says that intention is not reducible to a complex of belief and goals. The reason is that intention has specific properties with respect to action that do not have other mental states as epistemic states or motivational states.

As it has been explained above, it is very difficult to formalize the *intention with* and the *intentional action* (in the sense of: to act intentionally). In the area of computer science, the great majority of works (if not all of them) is about the third type of intention: the future-directed intention.

Future-directed intentions. Until now, we have only spoken about intentional action and with-intention; but what about expression of intention (that is, future directed intention [20], pure intention [12] or Searle's prior intention [137])? In the agent community, future-directed intention is certainly the most used kind of intention because it plays a major role by selecting among the goals (or the desires) of the agent those goals (or desires) that must be achieved or maintained. Roughly speaking there are two kinds of intention: intention *to maintain* something true (one believes that p is true and one intends that p remains true) and intention *to make* something true (where one believes that p is currently false and one wants to make it true).

One of the most influential work on future-directed intention is certainly Bratman's[20]. Following Casteñada [26], he defends the fact that future-directed intentions are inputs for practical reasoning[12] and planning whereas until now practical reasoning was just a belief-desire based reasoning where more or less importance is given to each of these mental state (see [40] for instance). For Bratman, future-directed intention is a special kind of motivational mental attitude because it is not just a *potential influencer of conduct* like desire for instance: it is "a *conduct-controlling* pro-attitude" [22,

[11]Note that the self-referential aspect of this definition entails that such a definition is particularly hard to formalize and, as far as we know, there is no logical formalization of the Searle's intention definition.

[12]Practical reasoning aims to define what to do in a particular situation.

p. 22]. In other words, it continuously drives and adapts our action policy and oblige us to plan its own realization, sometimes by the way of some other (sub)intentions.

In the multi-agent systems area, intention is always in the first place a future-directed intention (when an agent has not begun to act yet.) When the agent has begun to act, this intention becomes an intention in acting (a with-intention) even if the difference between those two types of intention is generally not explicitly made. It follows from this use of future-directed intentions that every action of agents is generally *considered* as an intentional action (in the sense that an agent cannot perform an action without having the corresponding future-directed intention).

At the same time, some simple examples are sufficient to convince us that future directed intentions are neither necessary nor sufficient (in a logical sense) for an action performance because we can perform actions without such kind of intention and we can have such an intention without intentionally performing the intentioned action. Such a point of view is supported by several well-known examples ([137, p. 84] and several famous examples in philosophy, such as Chisholm's unintentionally killed uncle [30], Davidson's climber [40, Essay 4] and the above Example 1.1, or Bennett's killer, etc.). For instance, suppose I currently satisfy my prior intention to write this dissertation, but I write this sentence although I did not have the prior intention to write it. In this example, future directed intention to write this sentence has not been a necessary condition for writing it. Conversely, I can intend to write this sentence but never write it (because I have changed my mind, or because this sentence was not correct/necessary/beautiful/etc.) and future-directed intention is not a sufficient condition here.

But if future-directed intention is neither a sufficient nor a necessary condition of an action, why do future directed intentions exist? As Bratman shows, a main reason is that it is convenient for humans: due to the fact that intention[13] has a strong inertia, humans are not continuously deliberating about the same point and this deliberation is made in advance (as far as possible). Moreover, once an intention has been adopted by a human, it forces him/her to elaborate a plan for satisfying it and to coordinate his/her actions. For instance, suppose I have a glass of water in my hand and I intend to drink it. Moreover, suppose now that somebody ask me for some water: I can take into account his/her request and intend now to drink only

[13]In this paragraph, *intention* refers to *future-directed intention*.

half the glass. Thus, future directed-intention allows us to determine how actions should be executed.

As part of a plan[14], future-directed intentions will generally persist in the future all along this plan even if they support revision when it is necessary. It means at least two things: first, there is a kind of commitment from the agent here and this commitment will be dropped only in some "serious" circumstances; second, adoption of new intentions is constrained by intentions that have already been adopted. (Generally, the new intention must not conflict with (the realization of) the old ones.) This intention inertia principle is a fundamental property of intention. For instance (see [20]), suppose that agent i intends to go to the dentist and that i believes that if it goes to the dentist then it will have a big pain in its teeth, does i intend to have a big pain in its teeth? For giving a response to this question, we just need to consider the following situation: suppose i intends to have a big pain in its teeth and suppose moreover that just before going to the dentist, agent i learns that by taking medicine it will not have any pain. Surely, agent i will prefer to take medicines and then, accepts to drop its intention to have a pain and to adopt the intention to take medicines. But it is contradictory with the principle saying that new intention are constrained by old ones and then, we cannot say that agent i intends to have a pain; it is just a side-effect of another intention (to go to the dentist).[15] Finally, this principle gives us a criteria for intention identification.

In Cohen & Levesque's works, intention is formalized in a reductionist way from belief and goal (that is, as a complex of beliefs and goals). The resulting modal operator of intention is defined in a non normal modal logic. (In particular, the K axiom and the necessitation rule does not hold for this operator.) Note that Cohen & Levesque claim [32]that they formalize intention following Bratman's work whereas in Bratman's view, intention is a concept that is non reducible to such a complex of belief and goals. A couple of years later, Rao & Georgeff defined intention with the help of a normal modal operator of intention [122, 123]. The formal framework is based on temporal logic but some properties are counterintuitive. For instance, an

[14]Here, a plan is not a kind of recipe or of template that we use when we need it, but rather it describes what we want now to do in the future. It is typically partial and has a hierarchical structure.

[15]Moreover, Bratman demonstrates that even if the agent *chooses* action β as being in the same *package* as action α, it does not intends it. In other words, this inertia principle is a property of rational choice but not of intention.

agent intends every tautology. But this property could not be desirable even if intention was here a maintenance intention because a tautology remains true whatever the intention of the agent. Moreover, together with the rationality of belief (D axiom) this framework entails unwanted intention (see the above dentist problem). As we will show in the next chapter, some other problems are related to the fact that intention is formalized with the help of a normal modal operator.

1.2.4 About normative mental states

Finally, normative mental states are internalized normative values. Normative value are every value that should be true in the world or believed as such (for one reason or another). Normative values include obligation, permissions and interdictions with respect to the law, traditions, religions, moral, social rules, etc.

Internalized values correspond to values that an agent forces itself to respect, whereas other normative values (to which the agent does not adhere) can be viewed just as beliefs about norms. Thus, we cannot say that these other normative values are mental states. For instance, a professional killer can believe that killing is forbidden and believes at the same time that she/he has no reason for observing this norm. In this case this norm has not been internalized and we cannot say that the killer has a normative mental state. On the contrary, if this norm has been internalized by an agent, then this agent will not kill anybody because from its point of view "it is bad/forbidden" and because it aims at respecting this norm.

It may happen that this aim is contradictory with respect to another value or another desire. For instance, I believe it is forbidden to run a stoplight but an ambulance behind me wants to go in front of me and thus I choose to go through the red light. In this case, to adopt a goal is make a choice between desire(s) or moral value(s) and it can happen that this choice violates a moral value. It does not mean that the moral value was not internalized, but that we have a hierarchical organization between our desires and moral values: some desires and moral values are more important than others. In such a case, violating a moral value has always a "moral cost" (more ou less important).

1.2.5 About action

What could be an agent without any relationship both with its physical environment and with other agents? It could neither perceive any new piece of information from this environment nor interact with (or act on) it. It just will be a misanthropic agent cut off from the world, a kind of useless bacterium that is unable both to eat and to grow!

But in case of agent systems such as they are designed in computer sciences, a theory of action is a fundamental part of agents capabilities: mental states and reasoning on the one hand, and action management on the other hand, are the two faces of a same coin. Action allows agents to (try to) change the world in accordance with their intentions. By "the world" we mean here everything outside or inside its mental states. More precisely, an action can change the physical world but also the mental states of agents (the author of this action included). For instance, by tossing a coin, agent i changes:

- the physical world (for instance, the coin is not at the same place),
- the mental states of the agents who were present during the toss action (for instance, they all believe that agent i has tossed a coin),
- and the mental states of agent i itself (for instance, it learns that the result is heads).

In the previous section, we have said that it is very difficult to define the concept of intention because, due to the fact that there are three concepts of intention, it is hard to find a definition that takes into account each kind of intention. The second reason, that we explain in the following of this section, is that relations between intention and action are very complex.

What is an action? From a philosophical point of view, one of the most important properties of actions is the relation between an action and its author, the underlying hypothesis being that, contrarily to events, agents *have something to do with action*. It is usual to try to distinguish what happens to an agent (for instance: Lila grows, Tom has been struck by lightning, somebody has sent a ball on Kenzo's head, etc.) from what this agent does (she eats an apple, he changes gear, they are sneezing, etc.). For Davidson "there is a fairly definite subclass of events which are actions" [40,

Essay 3] and an event is an action when there exists a causal relation (called *agency*) between an agent and this event. It is also not easy to define what is agency for at least two reasons.

First, we may do something without performing an action. For instance, when Lila upsets her cup of tea because Tom shakes her arm, Lila has upset her cup of tea (it is what she did) but in some sense, she is a victim: her body has upset (in an unintentional manner) the cup but somebody else has moved her body! Thus, we clearly cannot say here that she has performed the action of upsetting her cup of tea and here to upset the cup of tea is not an action at all: it was just, from the Lila's point of view, an *event*.

A second reason is that sometimes we believe to perform an action α while we are in fact performing an action β. We say thus that we have performed an *unintentional action* (that is β). As noted by Searle [137, p. 82] this kind of action is possible only because another action, identical to the first action, is performed simultaneously. For instance, let us suppose that agent i believes intentionally to upset a cup of tea.[16] But suppose now that in fact it is a cup of coffee (but agent i believes it is a cup of tea): as agent i intended to upset a cup of tea, what is the status of *to upset a cup of coffee* here? In some sort, we can say that a part of this event is intentional. This part is related to the action "to upset a cup" and follows from the performance of the identical action *to upset a cup of tea*. Thus, *to upset a cup of coffee* is an action. But as i did not intend to upset a cup of coffee but a cup of tea, *to upset a cup of coffee* is an unintentional action (although a part of this action is of course[17] intentional).

Let us note that there is a subtle difference between unintentional actions and *unconscious actions*. We sometimes perform some actions without being aware that we perform them (and in this case we do not believe that we are performing another action). For instance, when we drive a car, we can both discuss and shift gear. In this case, we are not conscious to perform the action to shift gear. Nevertheless, this action is intentional although it is not performed consciously: it wants to satisfy an intentional subgoal of a more general plan (to drive a car).

Finally, one the one hand we have unintentional actions and on the other hand we have things that we do unintentionally that are not actions at all. Thus, as noted by Davidson, if intentionality (in actions) entails agency

[16]This example is adapted from [40, Essay 3].

[17]We say "of course" because "there are no actions without intentions" [137, p. 82].

the converse is not true: We can perform actions unintentionally. Thus, agency distinguishes unintentional actions from events. Following Searle, this "agency component" of unintentional actions is related to the action that we have intentionally performed. The different cases are summarized in Table 1.1. As intentionality entails agency, the case where there is intention-

Table 1.1: Action *versus* event

Event properties		**Agency**	
		Yes	**No**
Intentionality	**Yes**	Action	$\emptyset$
	No	Action	Event

ality but no agency is impossible (which is represented by the $\emptyset$ symbol).

Representation of actions. In AI, there are roughly two ways for representing actions. The first one considers actions as a explicit label (in the object language) associated both to a set of preconditions and to a set of effects. In AI, this is typically the way taken by Harel's dynamic logic [70]: when preconditions are fulfilled the action is executable; when this action is executed, then the effects happen. Thus, actions have no truth value: actions are not propositions (viewed as sets of possible worlds) but relations between possible worlds. These operators speak about what will be true just after the execution of the action or about what was true just before this execution. In this view, we can thus only observe the world just after or just before the execution of an action, but never during this execution.

According to the second way, actions are not syntactically explicit (there is no action entity at all) and are represented by the state of affairs resulting from the execution of the action we are speaking about. This resulting state is formalized by a STIT operator (for Seeing To It that) [15]. For instance, the action of opening a door is represented by the fact that an agent sees to it that the door is open. This kind of operators is particularly well adapted for multi-agent systems with joint actions.

Speech acts. Since our Master thesis, we have worked on speech act theory. This is a philosophical theory where utterances are viewed as particular

actions [133]. There are several kinds of speech acts but the most known is the *illocutionary act*. (Often, people use the term "speech act" instead of "illocutionary act".) An illocutionary act describes what use of a given utterance we want to make by specifying an *illocutionary force* and a *propositional content*. There exist five *classes of illocutionary forces* [136]: 1) assertive (to commit the speaker to the truth of something), 2) directive (to attempt that the hearer does something), 3) commissive (to commit the speaker to do something), 4) declarative (to do something merely in virtue of the fact that one has declared that this thing is being done), and 5) expressive (to express psychological states about the propositional content). The propositional content represents the logical part of utterances, that is, the part that can be true or false. The author of the act is the *speaker* and the addressee of the act is the *hearer*.

Let us consider the following utterances (adapted from [133, p. 22]):

- Sam is a smoker
- Sam, be a smoker!
- [I promise that] Sam will be a smoker
- [I declare that] Sam is a smoker!
- I regret that Sam is a smoker

The propositional content of each sentence is the same ("Sam is a smoker") but the illocutionary force is different and its type is respectively: assertive, directive, commissive, declarative and expressive.

Illocutionary acts are generally associated to six conditions [156, Chapter 4]: the illocutionary point, the condition on the propositional content, the sincerity condition, the preparatory condition, the degree of strength, and the mode of achievement (see example in Tab. 1.2). It is important to note that there is one and only one illocutionary point by class of illocutionary act. Moreover, the illocutionary point is determined in part by the direction of fit between the world and the words. For instance, an assertive act is a description of the world (the direction of fit is from the word to the world) whereas both directive acts and commissive acts describe the world as it should be in the future (direction of fit from the world to the words). Declarative acts have both directions of fit whereas expressive have no direction of fit (they describe a psychological state, not the real world).

Table 1.2: Conditions (following [156]) of the illocutionary act from agent i to agent j realized by the utterance of "Could you open the window, please?"

Condition	**Example**
illocutionary point	agent i ask agent j to do something (class of requests)
propositional content	the sentence describes a future-directed action of agent j by using future tense
sincerity	i really wants that the window be opened by j
preparatory	i believes that j is able to open the window
degree of strength (of the sincerity condition)	neutral (agent i does not especially insist on its request)
mode of achievement (of the illocutionary point)	an option of refusal is given to j by i

These conditions can be fulfilled following different ways [156, p. 129–134]. We say that an illocutionary act is *successfully performed* (or *successful*) if and only if, by performing this act, the author of the act *expresses* its sincerity condition and its preparatory condition, and the other conditions are true.

An illocutionary act is *non-defectively performed* (or *non-defective*) if and only if the speaker is really sincere and the preparatory condition is really true (and the other conditions are also satisfied).

Finally, an illocutionary act is *satisfied* if and only if there is a fit between words and world.

As we will show in the following, the non-defective condition is particularly important as soon as we want to catch the public aspect of a conversation (see Section 3.3 for more details).

1.3 Short history of cognitive systems

There has been an old history between logic and reasoning since Aristotle and we can say that logic has been developed with the aim to reason. In particular, modal logics has been used because the truth of modal formulas is not a function of the truth value of their subformulas, what is especially interesting in case of mental attitudes: we can believe the sky is blue even if the sky is not blue (we make a mistake). In other words, a formula rep-

resenting a belief may be true even if the object of this belief is false in the real world.

Some works in computer science (see works of Allen, Cohen and Perrault for instance) and some others in philosophy (see [137] and [20] for instance) have cumulated at the end of 80ies to the BDI logic (for belief, desire and intention) in Cohen and Levesque. In this logic, intention is defined in a non primitive manner from both belief and goal, and agent theory is viewed as a particular action theory [32].[18] This formal framework is also used for communicative acts representation [34]. We can say that these publications have been the year zero of BDI logics.[19]

These works have been followed by those of Rao and Georgeff: in a more rigorous framework (based on a temporal tree structure with an axiomatics and a semantics), intention is defined as a primitive concept (see [122, 64] for instance).

As follow-up works one may cite cite the LOgic of Rational Agent (LORA logic) of Wooldridge [160]. The aim is not only to define a BDI agent architecture, but also to define its temporal evolution. Sadek's works also directly builds on Cohen&Levesque works. It is a BDI agent architecture for a rational interaction. This work strongly influenced the FIPA agent communication language.[20]

Since the middle of 90ies, conceptual analysis of mental states have been well-known and new BDI languages appear not only for describing agent architectures, but also for implementing them. Situation calculus is used as a BDI language, and real programming languages are associated to it (see GOLOG or ConGolog for instance). These works are named today cognitive robotics (see the group of Levesque and Lespérance at Toronto for instance [74, 140]).

At the same time, norms have been added to BDI frameworks. An agent not only can believe or aim at something, but it may also have obligations as in the BOID architecture of van der Torre (where 'O' is for "obligation") for instance [23]. This is an important step: obligation can be seen either as an internalized concept or as (a belief on) an external constraint.

[18] It is one of the reasons for which speech acts theory [13, 133] has had a great success in computer science area: in this theory, speaking is viewed as performing a particular action, which facilitates formal union between physical actions and linguistic actions.

[19] Their article in *Artificial Intelligence* received the *AAMAS most influential paper award* in 2008.

[20] http://www.fipa.org/repository/aclspecs.html

Moreover, BDI systems are criticized in agent community because of the strong assumptions that are made during their formalization. It is particularly the case for the sincerity hypothesis: in the FIPA agent communication language for instance, an agent believes everything what other agents say because every agent assumes that others say only truths. (They are assumed to be both sincere and competent about everything.)

Thus, at the end of 90ies, some alternative works focus on external concepts where interactions between agents could be described without any hypothesis on their mental states. The most influential works in this area are those about social commitment where every speech act of an agent is viewed as a particular commitment on the content of this act. For instance, when agent i says p is true, then it is committed on p: henceforth, agent i can neither says p is false nor act as if p is false. (Before that, it must say explicitly that p is false.) See for instance works of Singh [141] or of Colombetti [35]. Nevertheless, this approach make their own assumptions (public performance of actions for instance) and they do not explain how these new concepts are related to mental states. Often, these concepts are represented as primitive operators and relation with other concepts such as obligations, violations, beliefs, etc. are not described. These approaches are thus at the border of BDI approaches because they do not require any mental state.

At the same time, researchers realised that traditional AI problems also occur in BDI frameworks: frame problem (how formalize in an economic manner the fact that a lot of pieces of information does not change after the performance of an action?) ; qualification problem (what are all the preconditions of actions?) ; ramification problem (how describe the effects of effects on both the agent's environment and the agent itself?). With BDI frameworks, some problems must henceforth be solved: belief revision, update, expansion, contraction (see [154] for instance). More recently, these last problems are the core of dynamic epistemic logics (see [155] for instance).

Finally, some agent platforms have been developed as AgentSpeak by Rao, Jason by Hübner & Bordini or 2APL by Dastani. These platforms allow implementations of multi-agent systems without using until now the entire expressive power of BDI logics. In particular, these platforms do not have a complete set of boolean operators and do not allow for higher-order belief.

BDI concepts have also been used in other areas. It is the case for argumentation for instance, where Amgoud uses argumentation to generate both of desires and of plans in an autonomous agent [7].

1.4 Conclusions

We have presented above our view of cognitive agents. These are agents having mental representations of the world (with the help of belief or knowledge), and having motivational states (with the help of desires, goals, or intentions). Sometimes, cognitive agents can also have normative states. They are able to reason and can deduce new pieces of information from existing mental states. They can also perform actions in the aim of transforming the world with respect their motivational states.

These mental states and these actions can have several properties depending on what we want to do. For instance, agents can be sincere or not, they can be rational or not, etc.. They also can be defined in several ways, depending on what particular properties we want to focus. For instance, intention can be defined in a primitive manner or as a complex of belief and desire; action can be explicit or not in the logical language used by the agent, etc.

Clearly, agent architectures based on mental states allow explanations of the agent's behavior: when an agent performs an action, we are always able to explain which mental states have motivated the performance of this action. But every coin has two sides, and such agent reasoning has clearly a cost from a computational point of view. This point explains that implemented models of BDI architectures are often not really based on mental states and on reasoning on these mental states.

More generally, BDI architectures have been criticized for several years, because they induce several hypothesis on mental states (see [142] for instance). It did mainly concern sincerity (but not only) and more generally the fact that mental state are private states (in the sense that these states are only accessible to the agent having them). Such systems oblige agents to make hypotheses about the internal functioning of others, which can be problematical in case of a system with heterogeneous agents. But it has to be noted that this criticize concerns only logical or computational models, not philosophical view of mental states and this point requires some explanations. The strongest argument is that humans have only (private) mental states but successfully communicate each other. Let us illustrate that point with the help of an example.

Suppose that agent i says to agent j that p is true. If agent i is sincere, we have to believe that i believes that p is true (see [133, 156] for instance). And if we suppose that agent i is competent about p we believe that i is right

in believing that p is true. At the same time, if i is sincere, we can suppose that i wants to communicate to j the fact that p is true (we suppose that i does not make here a phatic act). From these facts, j can finally deduce that j also believes p is true. If in original Sadek's theory [130] these different steps were distinguished each other, it is not the case in the well-know agent architecture FIPA [51], although this architecture has been based on Sadek's works. Thus, it follows from this framework that, as soon as agent i informs j that p is true, then j believes p, which is clearly a too strong hypothesis. But we think that as soon as an architecture respects the two steps process of belief adoption (sincerity and competence) this architecture remains very close to human process (even if humans certainly often reason with the help of something like default rules [126]).

Another criticism is that, due to the fact that private mental states can differ from public attitudes (see Section 3.2 for more details), we need some other concepts to catch public norms, public commitments, group attitudes, etc. In the following chapter of this work, we argue here that we can deal with *expressed* mental states (see above) and other social attitudes (see Section 3.2). Be that as it may, we think that every social attitude (such as social commitment for instance) is necessarily connected to private mental states because our social behavior depends on our mental states. In social approaches such as commitment based architectures [35], social concepts have explicitly been separated from mental states. We hope that a part of our work has contributed to their reconciliation.

Chapter 2

Overview

2.1 Research themes

Our works are in the area of knowledge representation and reasoning. In this area, we have been concerned both by mental states and by speech acts theory [133] since the beginning of our works. It concerns not only the representation of these mental states but also their dynamics (through the execution of physical action or of linguistic actions) and cognitive agents have always been a natural way of investigation.

The cognitive agents area can be grasped at three levels:

1. the theoretical foundations (how to characterize cognitive agents?);
2. the agent communicative languages (ACL; how to characterize communication between agents?);
3. the implemented agents (applicative issues).

Until now, our works have been mainly concerned by the two first levels (even if we have also done some works at the third level). Each time, our studies have been formalized with modal logics which are the formal tools of our team.

The first level collects the main part of our works. What are the both sufficient and necessary basic concepts for designing a cognitive agent? What are their properties? (For instance, are the agent's beliefs rational? Is it sincere? etc.) What are the relations between each concept of the designed system? (For instance: if an agent does not know if p is true or false, does it

aim to know if p is true or false?) What are their relationships with human cognition? (For instance, is rationality an intuitive property with respect to human cognition? If not, which hypothesis this property entails?) etc. By "basic concepts" we mean concepts that correspond to basic cognitive components of cognition. For instance, the belief or the knowledge allows to represent the world from the point of view of an agent. Choice, goal, desire, etc. are several (similar but different) concepts for catching how an agent wants the world to evolve. Thus, we have studied some complex concepts built from basic ones, such as intention "à la Cohen and Levesque" [32], trust, responsibility, social concepts (group belief, group acceptance, common belief, institutions, delegation, reliance, etc.), emotions, etc. Each time, our study has also concerned the properties of theses concepts, and their relations with other (basic or complex) concepts.

From a technical point of view, basic concepts should intuitively be formalized by a primitive modal operator while complex concepts should be formalized by a formula using these primitive modal operators. But depending on our needs, complex concepts are sometimes defined with the help of a basic operator. For instance, we have formalized intention as a complex operator but also "à la Rao and Georgeff" [122] for which intention should be formalized as a primitive modal operator.

As we are concerned by autonomous agents here, we have always preferred working with belief rather than knowledge. From an epistemological point of view, knowledge is often viewed as true belief: agent i believes p is true and p is true in the real world. But for technical or purely theoretical reasons, we have sometimes used the concept of knowledge.

At the second level (the level of ACLs) we have been especially concerned by speech act theory. Since our master thesis, we have been convinced that speech act theory is able to represent communication (in the general sense), even if the original theory only concerns verbal communication. This part of our works concerns both theoretical aspects of speech act theory (such as indirect speech acts or non-literal speech acts for instance)[1] and practical

[1] A part of these works has been managed together with neuropsycholinguistical studies. If a literal utterance can be understood *via* normal cognitive mechanisms, hearers must recognize that the non literal utterance is deviant before determining his meaning. The traditional point of view, suggests that it would be more difficult to process non literal utterances than their literal counterpart. A modified version of this model, named "multiple-meaning model", suggests that comprehension of non literal meaning involves a simultaneous processing of literal and non literal meaning, but not a sequential processing.

aspects (such as: how to formalize some speech acts in the aim encoding a protocol of communication). From our point of view, ACLs and protocols on the one hand and mental states on the other hand are strongly related because speech act conditions depend on mental states before the performance of these acts. (See [56] for instance.)

Our works have also concerned the third level of the cognitive agents area. We have been interested in theorem proving (how to define and how to implement a generic theorem prover named Lotrec). The aim of Lotrec is not efficiency but to be generic: almost all modal logics can be implemented with Lotrec. More recently, we have contributed to develop with Lotrec a logical model of emotions. It is essential to have automated theorem proving methods for the logics under study because it allows both to play with the implemented logic and to test it with different kinds of queries. We have also been interested by cognitive architectures such as ACT-R.[2] A cognitive architecture aims to explain how central cognition (the core of the cognition where reasoning happens) works. For instance, how do we retrieve information from declarative memory? How do we execute mental action for reasoning? etc. Such architectures can provide some ideas for understanding dynamics of beliefs.

2.2 History and not detailed contributions

In the following, we speak about our contribution to cognitive agents from a chronological point of view. The aim is to show their main thread. Thus, we show how the articles presented in the two next chapters are a part of a larger plan of this contribution. We give a more detailed view of some important points that are not detailed in the next chapters.

2.2.1 PhD thesis (1995–1999)

Our Master Thesis was about speech acts theory. We mainly studied the works of Austin, Searle and Vanderveken. It was the first time we heard about speech acts theory but retrospectively, speech act theory has had a strong influence on our work and has been at the origin of our PhD thesis.

Our aim was to formalize these subtle differences. (See [28] for more details for instance.)

[2]We worked on ACT-R during one year of post-doctoral studies in the CLLE-LTC group of cognitive psychology (UMR 5263).

During our PhD Thesis (oct. 1996–nov. 1999) we were working about man-machine cooperative dialogue. More precisely, we were interested by belief change of interacting agents. (The title of this PhD was: "Rational interaction and belief change in dialogues: a topic-based logic".) It was the theme of a team project together with France Telecom R&D (which is called Orange R&D now). From a technical point of view, the aim was to define a logical framework for both mental attitudes (such as belief and intention for instance) and (linguistic) actions. During a conversation, an agent perceives new pieces of information that can be either consistent or contradictory with its current mental states. In the latter case, the agent must change its mental state; but how to describe that process? And how to represent this *new piece of information*? What are the links between it and the current mental states of the agent?

During theses works, we have been interested by: BDI agent architectures and formal theories of rational interaction [32, 129, 122]; belief change [5, 91] and modal logics [29, 119]. We have also studied topics theory (or themes theory). Except topics, all these themes[3] have represented a great part of our work since our thesis and more details are given in the next chapter. Thus, we just give some details about topics here.

In FIPA agent communication language (FIPA-ACL for short), agents are considered sincere and competent about anything. But this is a strong assumption because when an agent i says that p is true to an agent j, thus agent j necessarily believes that p is true. (It believes everything agent i says.) It is not the case in the Sadek's original agent theory where belief adoption depends on both sincerity and competence of agent i about p.

In the aim to relax this assumption, during our PhD we introduced an original association between topics and logical components of our framework:

- topics associated to a formula represent something like: "what is this formula about?" and such a set of topics is called *the subject* of this formula;
- topics associated to an agent represent the competence area of this agent and the set of such topics is called the competence (area) of this agent.

[3]Together with the dynamics of intention from 2002 to 2004, leading to a revisited version [83] of the most influential BDI logic, that is, the Cohen & Levesque's BDI logic.

- topics associated to an action represent the beliefs impacted by this action and they are called the *scope* of the action.

For instance, the atomic formula *weatherIsNice* could be about the weather and $Bel_i\, Intend_j\, skyIsBlue$ is about the color of the sky. Topics associated to agents describe the (informational) competence areas of this agents. For instance, agent i could be competent about the color of the sky but not about the rules of poker. Finally, when agent i informs j that the sky is blue (we say here that agent i performs an illocutionary act towards agent j), agent i intends that j believes that the sky is blue, and the effect of this information act is that agent j believes that agent i has this intention. In other words, the topics associated with this speech act include at least those about a mental attitude of i about a mental attitude of j about the color of the sky.

Note that from a technical point of view, topics sets are only defined for atomic formulas and a meta-theory of topics allows to compute the set of other (complex) formulas.

In our framework, some logical axioms depend on constraints between the different sets of topics. For instance, if agent i believes φ then p is true only if topics associated to φ are a subset of the topics representing the competence area of agent i. Formally:

$$Bel_i\, \varphi \rightarrow \varphi \qquad \text{if } \mathsf{Subject}(\varphi) \subseteq \mathsf{Competence}(i)$$

where $\mathsf{Subject}$ maps formulas to a topics set and where $\mathsf{Competence}$ maps agents to a topics set.[4] We consider that the above principle is a logical axiom and it entails for instance that $Bel_j\, Bel_i\, \varphi \rightarrow Bel_j\, \varphi$ which is read: if agent i is competent about φ then, if agent j believes that agent i believes φ then agent j believes φ.

A solution of similar nature has been given for the traditional frame problem: a formula is preserved after the performance of an action if this action and the formula do not share any topic:

$$Done_\alpha\, \varphi \rightarrow \varphi \qquad \text{if } \mathsf{Subject}(\varphi) \cap \mathsf{Scope}(\alpha) = \emptyset$$

[4]In a similar manner, a set of topic can be associated to the sincerity areas of an agent. Thus, when an agent j believes that an agent i intends that agent j believes that p is true (formalized as $Bel_j\, Intend_i\, Bel_j\, p$), we can deduce that agent j believes that agent i believes that p is true (formalized as $Bel_j\, Bel_i\, p$). We can thus apply the above competence axiom.

where $\mathsf{Scope}(\alpha)$ is the set of topics associated to α and where φ is not of the type $Done_{\alpha'}\,\varphi'$. This last restriction is due to technical reasons: such formulas cannot be preserved because they read "the action α has just been performed" and thus, it does not make any sense to preserve them. The axiom above is read: if the action α does not influence the truth value of p then, if p was true just before the performance of α then p is still true after this performance.

Moreover, we have shown that topics can be viewed as a meta-linguistic implementation of dependence in the sense of [27] and they can be rewritten as follows:

$$\begin{array}{lll} Bel_i\, p \rightarrow p & \text{if} & p \overset{c}{\rightsquigarrow} i \\ Done_\alpha\, \varphi \rightarrow \varphi & \text{if} & \alpha \not\rightsquigarrow \varphi \end{array}$$

where

$$p \overset{c}{\rightsquigarrow} i \overset{def}{=} \mathsf{Subject}(p) \subseteq \mathsf{Competence}(i)$$
$$\alpha \rightsquigarrow \varphi \overset{def}{=} \mathsf{Topics}(\alpha) \cap \mathsf{Topics}(\varphi) \neq \emptyset$$

and we note $\alpha \not\rightsquigarrow \varphi$ when this intersection is empty.

These axioms have been integrated into a BDI logical framework. The first results have been published in [50] but just assertive illocutionary acts were analyzed. This limit has subsequently been removed by introducing the concept of *contextual topic*, that is, a topic together with a context related to a (sequence of) mental state(s). For instance, the topic of $Bel_i\, Intend_j\, skyIsBlue$ is now about a mental attitude of agent i about a mental attitude of agent j about the color of the sky (rather than just about the color of the sky). This subtle difference has allowed to distinguish the expression of the speaker's beliefs (assertive act) from the expression of the speaker's intention that the hearer do something. It is an extension of Epstein's work on topics [47]: topics are not only associated with atomic formulas but also with modalities.

These results have been published in [77, 104, 79, 101]. We have shown that in a conversation between two agents the length of the context may have any value but two levels are often sufficient: the mental attitude of agent i about those of agent j about those of agent k (about ...) is the same as the mental attitude of agent i about those of agent j. A complete version of these results has been published in *Journal of Semantics* [78].

2.2.2 Multidisciplinary contribution around language (1999–2008)

Indirect illocutionary acts (1999–2006). In our PhD thesis we were interested only in literal communication. When we are speaking, sometimes we make literal illocutionary acts (by saying p, the speaker wants to say exactly p) and sometimes we make non-literal illocutionary acts (by saying p, the speaker wants to say q). Particular non-literal speech acts are indirect speech acts: in this case, q implies (in an illocutionary manner) p. In other words, when agent i says indirectly q by saying p, it wants mainly to communicate that q but also communicates that p. We have worked on indirect illocutionary acts together with linguists and neuroscientists.[5]

The utterance "Pass me the salt" for instance is a direct request where the *meaning of the utterance* fits with the *meaning of the speaker* which is, following Searle terminology, what the speaker wants to say [133]. But in "Can you pass me the salt?" we can easily imagine a context where the meaning of the utterance and the meaning of the speaker do not fit, that is, a context where the former is a yes-no question whereas the latter is a request that means "Pass me the salt". In this case we say that the request has been indirectly performed by the way of a yes-no question. Note that the property of an utterance to realize an indirect speech act is cancelable: in the above example, we can imagine that "Can you pass me the salt?" is just a question about the capacity of the hearer to pass the salt (but the speaker does not want that the hearer gives him/her the salt). This is a property of every non-literal act.

In fact, we often use such indirect illocutionary acts because these acts offer an option of refusal to the addressee of the (indirect) request, which is socially more acceptable. We could think that there are a lot of manners to perform an illocutionary act indirectly but this is not the case: we perform an indirect illocutionary act by making an assertion or by asking a question about the preparatory condition or about the sincerity condition of the corresponding direct act (see the Virbel's article [158] for more details). For instance, "Can you pass me the salt?" is a question about the preparatory condition of the illocutionary act realized by the utterance of "Pass me the salt". (This preparatory condition is here that the speaker believes that the hearer —the addressee of its request— can pass him the salt.) These results

[5] We have made one year of post-doctoral studies in a laboratory of cognitive psychology.

have been published in an international workshop on dialogue [84].

Jean-Luc Nespoulous and Maud Champagne (who are neuroscientists) have made experiments on humans having right hemispheric brain damages about indirect illocutionary acts. The aim was to exhibit the underlying cognitive process of indirect acts. More precisely, we wanted to answer the technical aspect of following question: is it necessary to understand the literal act for understanding the indirect act? Neuroscientists have made experiments on both healthy persons and persons with right hemispheric damages, exploiting that indirect acts processing capacity is located in the right cerebral hemisphere. They have measured the response time in both cases (that were similar, showing that indirect speech acts understanding does not necessarily depend on the direct one). Depending on whether we answer *yes* or *no* to the above question, the formalization of indirect acts in a theory of action is different. We have formalized the interpretation of indirect acts in several publications (see [28] for instance).

We have also worked on indirect acts together with Eric Raufaste, a cognitive psychologist, in the aim to analyse the understanding process of utterances from the point of view of the central cognition. The latter can be compared with central processing unit (CPU) of a computer: it is the reasoning center of our cognition (following Anderson [10, 11]). We have both formalized and implemented with ACT-R [98] a basic model of understanding based on a learning process. The first results have been published in [103].

Three years later, following our contribution (together with our Master student Raphaël Saban) about the concepts of good and of bad, we have defined a new model of indirect acts inference in a French journal of psychology [102].

Post-doctoral position (2002) and afterwards (2003). In 2002 we were at the Work and Cognition Laboratory LTC (today, the CLLE-LTC laboratory, UMR 5263) to work on the cognitive architecture ACT-R which is the implemented theory of Anderson. Following this theory, each piece of information (that is called a "chunk") is associated with a numerical value: its activation level. An agent may be aware of a chunk if and only if the activation level of this chunk is above some given threshold. The chunks are linked with the help of a semantic network. When the central cognition attempts to retrieve some chunk, some quantity of activation is added to this chunk and to every chunk that is linked with this chunk. (This process is

called the activation spreading.) This is the subsymbolic level of ACT-R. At its symbolic level, ACT-R is able to reason about every chunk having an activation level over the threshold. At this level, ACT-R can retrieve some chunk from the declarative memory (the set of all chunks) or from some buffer (such as visual buffer, goal buffer, action buffer, etc.).

With the help of ACT-R we have modeled the Stroop effect experiment. Some character strings naming colors are printed on a screen and the color used for printing can be either the same as the color named by the character string (congruent item) or another color (incongruent item). For instance, the word "green" is written in blue, "red" is written in red, "blue" is written in yellow, etc. After a learning step, some items are presented to human subjects which must identify the color given by the character strings. We measure both the response time and its correctness for each subject.

We can observe three facts: the mistakes of the same subject are less and less frequent; the responses of subjects are faster and faster; the response time for incongruent items are longer than those for congruent items. Our model should have a very close behavior (with respect both response time and correctness) to experimental data.

A publication about an ACT-R model must demonstrate that the new model makes at least all what the old models do. We did not have enough time to complete this task and this work has not been published (even if it has allowed us to know both the central cognition functioning and ACT-R). Later, we have worked with ACT-R again about indirect acts.

Relation between language and action (2007–2008). Together with Pr. Alain Trognon (psychology department of Nancy 2 University) we have published two articles with a strong psychological content. It was a work about the Hanoi towers problem solved by children. A child was not able to move pieces alone and two children had have to discuss and move the pieces together. The aim was to analyse their conversation. Results have been published in [147] and in [148].

2.2.3 After PhD Thesis

After our thesis, belief change (how do beliefs evolve after the performance of an action) has become an important part of our work and we have contributed to develop some other solutions for characterizing mental states and their dynamics (see Section 3.1).

During this period, we were also interested in theorem proving. Thus, we contribute with several members of our team for developing Lotrec, a generic theorem proving. Our main contribution was to make the interface between the theoretical works of the LILaC group and the first implementation of Lotrec by David Fauthoux. This work has been published in [49] and it has certainly been our first contribution in a PhD student supervision. Later, we have published (with our colleagues) some results about a new implementation of Lotrec [55].

We have said in Section 1.4 that BDI architectures have been criticized (among others) for their internal states hypothesis. Starting from this criticism, we tried to characterize speech acts (thus, linguistic actions) as the *expression* of some mental states, that is, the set of propositions that are publicly true after the execution of this act. The point is that when an agent i says: "the sky is blue", i necessarily expresses the fact that i believes that the sky is blue (but, maybe, i does not really believe that the sky is blue). (This point is related to the success *versus* non-defective performance of speech acts that we have discussed in Section 1.2.5 for more details.) Here, nothing is supposed about mental states of the hearer. We just use our linguistic competence. Thus, the question was: "is it possible to characterize how pieces of information are grounded during a conversation?". Here "grounded" means the pieces of information following from the performance of speech acts during a conversation and considered as true by the group of agents participating in this conversation.

This was a first shift in our works: whereas our previous works was until now about single agents, we begun to work on group attitudes. We studied respectively grounding and acceptance. These two concepts are a kind of group belief. In the first case, grounding is represented by a non-reducible modal operator and it is supposed that group belief does not imply individual belief, and that the belief of a group may differ from the belief of its subgroups. Our aim was to represent the belief of a group during a dialogue, even if some people leave the group or join it. In the second case, acceptance represents what agents, *qua* members of a group, accept as true. Here, group acceptance implies individual acceptance but it does not imply individual belief (the private part of agents' mental states). Acceptance is also related to an institutional context. Thus, in a context a group can accept that p is true, and in another context the same group can reject the fact that p is true. Acceptance is the foundation of institutions and grounding is about what a group considers to be true. Grounding has also been used for formalizing

Walton & Krabbe's PPD_0 persuasion dialogues. (See Section 3.2 for more details about our works on social concepts.)

The second shift in our works has happened in 2006 with the concept of emotion. It is not necessary to demonstrate anymore that emotion is a part of our cognition. There is plenty of works about that subject in psychology and in philosophy. Thus, extending our mental states analysis by emotion has been both a natural and a necessary step. At the same time, we also worked on the formalization of trust which can be viewed as a particular belief.

2.3 Conclusions

We have shortly described in this chapter the works that have not been described in the next chapter (that is: mental states and their dynamics, social concepts, and emotion). It follows from that presentation that mental states have been at the core of our works since the beginning even if three different steps can roughly be made: single agent (study of mental state and their dynamics); group of agents (study of social concepts useful for a group description, and study of the structure and of the properties of the groups); emotion (how to describe some complex mental attitudes, what are their properties, and how these complex attitudes influence the behavior of agents).

It is important to note here that speech act theory has always been present in our works since our Master thesis but they are not presented in the following in a particular section because they are transverse: we have studied speech acts of single agents, of agent groups (social concepts), and the expression of emotions. The next chapter presents some work on each of these three steps.

Chapter 3

Summary of selected articles

3.1 Mental attitudes and their dynamics

As said above, mental states are at the core of our works and we think that they must be at the core of agent systems. Thus, in the following we present some works on mental states and action (that are necessary as soon as we want to study the dynamics of these mental states).

In Section 3.1.1 we speak both about an original non-reductionist view of intention and about cooperation principles (how could an agent adopts some beliefs or goals/intentions of other agents). In Section 3.1.2 we present an original work based on Cohen & Levesque's works: we have recast their framework in a simpler way using the resources of dynamic epistemic logics and we have shown a lot of interesting properties. Finally, in Section 3.1.3 we speak about belief change and misperception.

3.1.1 Non reductionist view of intention

In [80], our aim was to propose cooperation principles for rational agents based on a non reductionist view of intention (that is, intention is not defined from other operators such as in Cohen&Levesque's view [32]). Cooperation means here that if agent i has a goal, then agent j will adopt i's goal as long as this goal does not contradict its own goals. Similarly, if an agent asserts something, the addressee of this assertion will adopt the beliefs conveyed by this assertion if it believes that the speaker is competent about what it says. From an internal point of view, the beliefs of an agent are preserved after the execution of an action as long as it does not contradict its own beliefs. This

works aims to propose a simple mechanism both for goal adoption and belief adoption in the general area of multi-agent systems where agents have some tasks that must be fulfilled and where they can ask for help to other agents. Thus, it concerns every systems where agents are not in competitions with others.

Intention definition. The BDI framework of cognitive agents has been reduced to the well-known doxastic logic KD45 [87] and modal operators for future-directed intention such that $Intend_i\,\varphi$ reads "agent i intends that φ be true" where the intention must be achieved (*vs* preserved). Such a definition entails that if agent i intends that φ then this agent believes that φ is currently false. It is the meaning of the following first principle:

$$Intend_i\,\varphi \rightarrow Bel_i\,\neg\varphi \qquad (\text{Rel}_{\text{IntBel}}1)$$

Let us note that a weaker definition such as $Intend_i\,\varphi \rightarrow \neg Bel_i\,\varphi$ does not properly describe rational intention to make something true because in this case, agent i can both together imagine at least a world where φ is false (that is $\neg Bel_i\,\varphi$) and intends φ. Thus, if φ is false in every epistemic world, we come back to our definition above. But if there exists at least one epistemic world where φ is true, it means that agent i intends φ while it believes that it is possible that φ be already true. Thus, intention would not just concern intention to make something true here, but it would also concern intention to maintain something true. (This is a more general kind of intention including both achievement and maintenance intention.) A consequence of that definition is that $\neg BelIf_i\,\varphi \wedge Intend_i\,\varphi$ (agent i does not know if φ is true or not, and it intends that φ is true), where $BelIf_i\,\varphi$ abbreviates $Bel_i\,\varphi \vee Bel_i\,\neg\varphi$, is contradictory. In this case, it means that if agent i does not know if φ is true or not, it should intend to know if φ is true or not. (That is: $Intend_i\,BelIf_i\,\varphi$.) For instance, suppose that agent i does not see anything at all and that it would like to turn the light off (noted $\neg light$). Thus, i does not know if the light is on or off ($\neg BelIf_i\,light$). But i cannot directly intend $\neg light$: it must first intend to know if the light is on or off ($Intend_i\,BelIf_i\,light$, which is consistent with the hypothesis $\neg BelIf_i\,light$).

Note that the above axiom entails (thanks to D axiom for belief) that

$$Bel_i\,\varphi \rightarrow \neg Intend_i\,\varphi$$

A second interesting principle is the link between intention and intention to believe:

$$Intend_i\,\varphi \rightarrow Intend_i\,Bel_i\,\varphi \qquad (\text{Rel}_{\text{IntBel}}3)$$

Thus, if agent i intends that φ is true, then it necessarily intends to believe that φ is true. In other words, agent i cannot intend to change the world without changing its mind about the new state of this world.

Finally, the last principle gives the converse link:

$$Intend_i\,Bel_i\,\varphi \wedge Bel_i\,\neg\varphi \rightarrow Intend_i\,\varphi \qquad (\text{Rel}_{\text{IntBel}}2)$$

It is read: if agent i intends to believe φ and i believes φ is false then i intends φ. In other words, we can intend to believe φ without intending that φ be true. $Intend_i\,Bel_i\,\varphi$ just entails $\neg Bel_i\,\varphi$ by ($\text{Rel}_{\text{IntBel}}1$) and the principles of our logic. Thus, only two cases are possible: i) $\neg Bel_i\,\varphi \wedge \neg Bel_i\,\neg\varphi$ (agent i does not know if φ is true or not); ii) $\neg Bel_i\,\varphi \wedge Bel_i\,\neg\varphi$ (agent i believes that φ is false). In the first case, it is still possible for φ to be true and thus agent i cannot directly intend that φ be true. (As we show in the following, the agent must first intend to know if φ is true or not before intending to make it true.) Thus, we cannot conclude in this case about i's intention about φ. In the second case, we obtain the above principle. Finally, when we want to believe something is true, it means that we can be in two different states: in the first case, we just have the intention to expand our believes whereas in the second case our *intention to believe* represents what we want to be true in the world, a doxastic prerequisite of a change of the world that we want to bring about.

Note that according several authors, "beliefs are involuntary, and are not normally subject to direct voluntary control" [152, 46, 21]. Note that the *intention to believe* that p is true does not entail that p will be a voluntary belief: the fact that we believe p does not depend on our voluntary but on facts of the world. In [80], intention is not defined *à la Cohen&Levesque* (that is, as a complex of goals and beliefs) but as a primitive concept *à la Bratman* formalized by modal operators (one for each agent i) in the object language of our logic. The principles followed by these operators are those of a classical modal logic rather than a normal modal logic. (In our previous works, we used a normal modal logic KD for formalizing intention (see [78] for instance) but as we show in the following, a lot of principles of this logic are counterintuitive.) A classical modal logic entails the validity neither of

the necessitation rule nor of the K axiom.[1] For instance, as we also have the rule of necessitation for belief, if we had the rule of necessitation for intention then for every valid formula φ we would have $Bel_i\,\varphi \wedge Intend_i\,\varphi$ that contradicts ($\text{Rel}_{\text{IntBel}}1$). These operators do not satisfy some other properties such as axiom of conjunction ($Intend_i\,\varphi \wedge Intend_i\,\psi \rightarrow Intend_i\,(\varphi \wedge \psi)$) because if agent i intends that two different things be separately true, it does not necessarily imply that agent i intends that these things be true both together. (A similar reason justifies that the contraposition of the axiom of conjunction, named axiom of monotony, be also rejected.) Finally, the only principle that holds for intention is the rule of equivalence

$$\frac{\varphi \leftrightarrow \psi}{Intend_i\,\varphi \leftrightarrow Intend_i\,\psi} \qquad (\text{RE}_{Intend_i})$$

saying that, if it is valid that two formulas φ and ψ are logically equivalent then it is valid that to intend the former is logically equivalent to intend the latter. Note that the modal logic of intention is thus as weak as possible because if we drop (RE_{Intend_i}) it would mean that intention is not a modality at all.

Action encoding. Joint to our BDI framework, we also need a formalization of some actions, in particular speech acts because they are fundamental for interactions between agents. We just use assertive acts and directive acts. One of the contributions of this article is to propose a formalization of directive acts with the help of assertive acts (see the paragraph about speech acts page 19 for more details about speech acts). Here, we just exploit our previous works about indirect acts (see page 32 for more details) by reformulating questions and requests with the help of assertions. Our aim is to propose a logic for linguistic actions that is as simple as possible. As said above, indirect acts are always cancelable but our aim is not here to formalize indirect speech act theory, but just to use this theory for proposing a convincing formalization of request and questions with the help of assertions.

Finally, when cognitive agents are able to understand speech acts expressing requests or questions they have to decide if they adopt the goal of the author's speech act or not. Thus, we need to elaborate cooperation principles in the aim to allow agents to adopt both beliefs and intentions of others.

[1] The rule of necessitation for intention says that "if φ is valid then agent i intends φ" and Axiom K says that "if agent i intends that φ entails ψ, then if agent i intends φ then agent i intends ψ".

Belief principles. Belief adoption is a part of cooperation behavior and it is defined with the help of a two steps process: one step for adoption of new beliefs and one step for preservation of some old beliefs. Belief adoption is similar of the process defined during our PhD except the fact that influence relations replace the topic functions. Formally:

$$Bel_i\,\varphi \rightarrow \varphi \qquad \text{if } i \overset{c}{\rightsquigarrow} \varphi \text{ and } \varphi \text{ is objective} \qquad (\text{Adopt}_{\text{Bel1}})$$

Note that it is a logical axiom and that necessitation rule applies. Thus, $Bel_j\,Bel_i\,\varphi \rightarrow Bel_j\,\varphi$ can be deduced from this axiom: if agent i believes that agent j believes φ then agent i believes φ, only if both agent j is competent about φ and φ is objective (that is, φ does contain any modal operator). This last restriction has been added here because it is not intuitive to suppose that agents may be competent on the mental states of the others. Moreover, if $\varphi(j)$ is a mental attitudes of j about an objective formula φ (for instance, $\varphi(j)$ is $Bel_j\,\varphi$) then $\varphi(j)$ is not an objective formula. If the above axiom would not be restricted to objective formulas it would entail for instance (after necessitation by Bel_j) that $Bel_j\,Bel_i\,\varphi(j) \rightarrow Bel_j\,\varphi(j)$, that is: if agent i is competent about some mental attitude of j about φ then agent j believes it has this mental attitude about φ as soon as it believes that agent i believes that agent j has this mental attitude. Clearly, it is not an acceptable principle. Thus, formulas that are not objective must be processed with specific principles (that are not described here).

The preservation principle is as follows:

$$\neg Bel_i\,\neg\varphi \rightarrow After_{\alpha(\varphi)}\,Bel_i\,\varphi \qquad (\text{Adopt}_{\text{Bel2}})$$

where $\alpha(\varphi)$ represents an assertive speech act having φ as propositional contents. From the point of view of epistemic worlds, this is a belief change principle as can be found in public announcement logic because after an assertion about φ the epistemic worlds where φ was false are no longer accessible. This is a valid principle even if the author of the speech act is not competent about φ. We can show that this principle entails $Done_{\alpha(\varphi)}\,Bel_i\,\neg\varphi \rightarrow Bel_i\,\neg\varphi$ that is, if an agent believes something is false before the assertion of its converse, then it still believes that this thing is false. In other words, when the beliefs of the agent and the propositional content of an assertion are inconsistent, the agent does not change its mind. Note that this axiom is slightly different from that has been defined during our PhD thesis (about our PhD thesis, see page 30) and the previous metalinguistic constraint (that does not allow the

preservation of $Bel_i\,\neg\varphi$ when $\alpha(\varphi)$ is performed) has been removed. Note also that, as $Bel_i\,\varphi \rightarrow \neg Bel_i\,\neg\varphi$, this new principle allows the preservation of beliefs that are consistent with what it is asserted. Nevertheless, this principle says nothing about preservation of other formulas.

Intention principles. We have proposed the following axiom:

$$(Bel_i\,Intend_j\,\varphi \wedge \neg Bel_i\,\varphi \wedge \neg Intend_i\,Bel_i\,\neg\varphi) \rightarrow Intend_i\,Bel_i\,\varphi \qquad (3.1)$$

that is read: if agent i believes that agent j intends that φ be true, and i does not believe dans φ is currently true, and i does not intend to believe that φ is false, then i intends to believe that φ is true. In other words, intention adoption is a two-steps process: when agent j intends φ then agent i intends to believe φ (if i believes it is possible that φ is true and if i does not intend to believes that φ is false); thus, thanks to ($\text{Rel}_{\text{IntBel}}2$), the intention to believe φ can be converted to the intention that φ be true (if agent i believes that φ is false).

Let us take an example. Suppose that agent j intends that the light be off in the office and that agent i believes that fact (that is, $Bel_i\,Intend_j\,p$ where p is read "the light is off in the office") and suppose that agent i does not know if the light is off or not (that is, $\neg Bel_i\,\neg p \wedge \neg Bel_i\,p$). Thus, if agent i does not intend to believe that the light is on ($\neg Intend_i\,Bel_i\,\neg p$), i intends to believe the light is off ($Intend_i\,Bel_i\,p$). Thus, agent i should perform some action in the aim to determine if the light is off in the office. Suppose that i go in the office and believes now that the light is on (thanks to ($\text{Adopt}_{\text{Bel1}}$)), that is $Bel_i\,\neg p$. Thus ($\text{Rel}_{\text{IntBel}}2$) applies and then agent i intends the light be off ($Intend_i\,p$).

The above framework allows us to deduce the following theorem:

$$(Bel_i\,Intend_j\,\varphi \wedge Bel_i\,\neg\varphi) \rightarrow Intend_i\,\varphi$$

that is read: if agent i believes that j intends φ and it believes also that φ is currently false, then it intends φ. In other words, agent i adopt intentions of others when it believes that the object of these intention is false. Note that $Bel_i\,\neg\varphi \rightarrow \neg Intend_i\,\neg\varphi$ and agent i cannot thus adopt intentions that are contradictory with those it already has.

Another interesting consequence is:

$$Bel_i\,Intend_j\,\varphi \rightarrow Intend_i\,Bel_j\,\varphi$$

that is read :if agent i believes that j intends φ then i intends that j believes φ. It is an interesting property because φ can already be true but agent j may ignore this fact. But suppose that agent i believes that φ is already true, then it can inform agent j about the truth value of φ.

3.1.2 Reductionist view of intention

The second article [83] proposes a quite different approach of intention. Here, intention is formalized *à la* Cohen & Levesque in a reductionist way. More precisely, intention is defined from persistent goals that are nothing but particular achievement goals. It is thus a major contribution to both a well-defined and a simplified version of Cohen & Levesque framework.

In the framework of this article, both belief operators and choice operators are defined in a KD45 logic. Choice is here viewed as something that one prefers to be true because it is a desire of oneself or because it is an internalized norm (see Section 1.2 for more details about these concepts). These choice operators are the same as the goal operators in C&L framework.

The two other components are action and time. Action is represented with the help of dynamic operator in a PDL style [70]. $[\alpha]\varphi$ is read "φ is true after every possible execution of the action α" and $\langle\alpha\rangle \stackrel{def}{=} \neg[\alpha]\neg\varphi$ is read "α is executable and φ will be true after some possible execution of α". Two different actions lead to the same world (linear time in the past and in the future). Thus, possible worlds are structured as histories (see Figure 3.1: w_0 is the real world, and w_0' w_0''... are epistemic worlds on linear epistemic histories). It is interesting to note that the set of epistemic possible worlds becomes a set of epistemic possible histories: thus, from the point of view of the beliefs of an agent, there are several possible futures (because the actions following α are not necessarily the same in each history).

Let us note that we make the hypothesis the beliefs of an agent about action occurrences are both sound and complete with respect to the action occurrences in the real world. That is, if an action happens in the real world, the agent believes that this action happens (the agent is aware of the action occurrences), and conversely if an agent believes that an action happens then this action really happens in the real world (the agent does not make a mistake).

Moreover, all actions are supposed here to be *uninformative*.[2] This hy-

[2]This is related to some previous works on that subject (see for instance [76, 75, 82]).

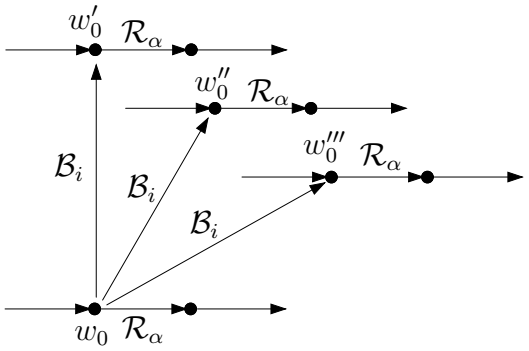

Figure 3.1: Belief, action and linear time

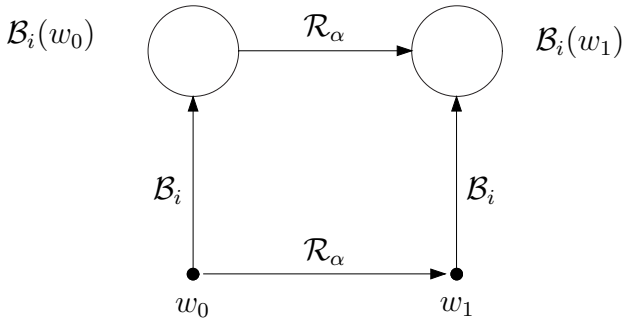

Figure 3.2: No learning and no forgetting hypothesis

pothesis allows us to model the belief preservation process as follows:

1. an agent does not learn anything from an action occurrence, and
2. this agent does not forget what it was believing after an action occurrence.

This can be illustrated on Figure 3.2. (1) means: if after the occurrence of α (that is, in w_1) the agent's beliefs are $\mathcal{B}_i(w_1)$, then necessarily, before the occurrence of α (that is, in w_0), agent i thought that after the occurrence of

More details and justifications are given in Section 3.1.3.

α its beliefs should be $\mathcal{B}_i(w_1)$. To ensure that w_1 exists it is required that in w_0 the action α is executable. (2) means: if in w_0 agent i believes that after the execution of α its beliefs will be described by $\mathcal{B}_i(w_1)$, then after the execution of α (that is, in world w_1) it will believes will be $\mathcal{B}_i(w_1)$. But it is intuitive only when the agent believes that the action α is executable. The hypothesis that actions are uninformative is necessary here because it entails that the execution of an action will not generate new beliefs that are contradictory with those that are preserved. Thus, (1) is formalized by the principle (NL_{Bel_i}) and (2) by (NF_{Bel_i}) and both allow the preservation of agents beliefs (under the hypothesis of uninformative actions).

Finally, temporal operators have relationships with actions. $\Box\varphi$ is read: "henceforth φ is true". The following principle :

$$\Box\varphi \rightarrow [\alpha]\varphi \qquad (\mathsf{INC}_{[\alpha]})$$

means that if something is henceforth true then this thing will true after every execution of any action α. Semantically, it means that worlds that are accessible with actions are a subset of those that are accessible with the temporal relation, but the converse is false. In other words, there exists at least one world that will be true in the future but that it is not accessible by the execution of an action. It allows an agent to believe that something will be true in the future while this agent does not know how this state of affairs will happen.

Thus, we are able to formalize intention now. Achievement goal is defined in a similar way as in [32]

$$AGoal_i\,\varphi \stackrel{def}{=} Choice_i\,\Diamond Bel_i\,\varphi \wedge \neg Bel_i\,\varphi \qquad (3.2)$$

but our epistemic condition is weaker: we only require that agent i does not believe that φ is true (that is, $\neg Bel_i\,\varphi$) whereas Cohen & Levesque require that agent i believes that φ is false (that is, $Bel_i\,\neg\varphi$). This condition seems us more natural here.[3]

An original aspect of our formalization of achievement goal is that it is also a persistent goal. Indeed, it is unclear for us what kind of achievement goal should not be a persistent goal. In any case, C&L never specify conditions

[3]Related to our discussion about the previous article, let us note that the choice is about a belief about φ whereas in C&L this is a choice about φ. (Following C&L, $AGoal_i\,\varphi \equiv Choice_i\,\Diamond\varphi \wedge Bel_i\,\neg\varphi$.) Thus, as we want to believe that φ will be true in the future, we just need that it possible that φ is false.

for which an agent should abandon such a goal. Thus, the following principle is a consequence of our framework:

$$(AGoal_i\,\varphi \wedge \neg Choice_i\,[\alpha]\bot) \rightarrow [\alpha](AGoal_i\,\varphi \vee Bel_i\,\varphi)$$

that is read: "If α is not an unwanted action, then after every execution of α either an achievement goal about φ is preserved or φ is believed as true".

Finally, intention that φ is true is defined as follows:

$$Intend_i\,\varphi \stackrel{def}{=} AGoal_i\,\varphi \wedge \neg Bel_i\,\Diamond Bel_i\,\varphi$$

Thus, agent i intends φ if and only if both φ is a persistent goal of i and i does not believe that it will believe φ sometimes in the future. In other words, agent i intends φ as long as, from the point of view of i's belief, there exists a history where henceforth agent i does not believe that φ is true.

Note that as soon as the agent i believes that in every possible history it will believe φ is true ($Bel_i\,\Diamond Bel_i\,\varphi$) then agent i does not intend φ anymore (that is, $Bel_i\,\Diamond Bel_i\,\varphi \rightarrow \neg Intend_i\,\varphi$). As the principles of our logic entail that $Bel_i\,\varphi \rightarrow Bel_i\,\Diamond Bel_i\,\varphi$, we can deduce that as soon as agent i believes that φ, agent i does not intend φ anymore (that is, $Bel_i\,\varphi \rightarrow \neg Intend_i\,\varphi$). Nevertheless, $\neg Bel_i\,\varphi \wedge Bel_i\,\Diamond Bel_i\,\varphi$ is satisfiable (we can say in this case that "agent i expects φ"), it is also the case that agent i does not intends φ anymore: this may appear very strange. Suppose both that I intend to turn the light on and that I believe after having pushed the button, the light will be on. (And suppose also that the action *to push the button* is executable.) As I believe that the light is currently off, I expect (in the sense defined above) that the light will be on in every possible future history. Thus, following our definition, I do not intend to turn the light on anymore!

In fact, we must here distinguish between different kinds of intentions. We recall that the above definition of intention is about a future directed intention. As it has already been explained (page 14) future directed intention is neither necessary nor sufficient for performing some action. It just helps us to organize and plan the future actions that we know we need to perform [22]. The (intentional) execution of an action requires an *intention in action*. That is, when we do something, we do it with the intention to do it here and now. We have formalized neither this kind of intention nor the change process from future directed intention to intention in action. We just suppose here that a future directed intention is dropped as soon as we have an executable sequence of actions leading to our achievement goal. (After

that, it is needed to have an intention in action for executing this sequence of actions.)

Note also that our definition is quite different from that of C&L because ours does not mention action at all. In [20] Bratman says that we need future directed intention for organizing what we have to do later; but in his opinion, it is not necessary that future directed intention defines the action sequence for fulfilling this intention: "plans are typically *partial*" [22, p. 19]. In contrast, C&L's definition of intention completely defines the action sequence that is necessary to perform for fulfilling this intention.

Finally, similarly to achievement goals, let us note that it can be proved that intentions are persistent. (It is an essential property of future-directed intention as shown by Bratman [20].) This property is formalized by the following principle (that can be derived from our framework):

$$(Intend_i\,\varphi \wedge \neg Choice_i\,[\alpha]\bot) \rightarrow [\alpha](Intend_i\,\varphi \vee Bel_i\,\Diamond Bel_i\,\varphi)$$

In other words, if agent i intends φ and if agent i does not prefer that action α is not executable, then after the execution of α either agent i still intends φ or it believes that, in each history that it believes as possible, it will believe φ at some time. Note that strong realism hypothesis entails that $\neg Choice_i\,\varphi \rightarrow \neg Bel_i\,\varphi$.

3.1.3 Sensing actions, belief change and misperception

The last article that closes this section focus on action [81]. We suppose that agents are in partial observable environment and that they can have a partial or an erroneous point of view about the real world. How can we describe the belief change in this context? What happens for instance when an action, that is not executable from the point of view of an agent, is executed? What is the belief state of this agent after the execution of this action? We propose in this article a belief change process including a revision operator. Actions are supposed to be uninformative.

In order to make things more simple, the BDI framework just includes belief. The logical framework is still based on PDL: $[\alpha]\varphi$ is read "the formula φ is true after every execution of the action α" and $\langle\alpha\rangle\varphi \equiv \neg[\alpha]\neg\varphi$ is read "α is executable and φ will be true after some possible execution of α".

A lot of actions have effects both on the physical world and on agents' mental states. We call the former the *ontic effects* of the action, and the

latter the *epistemic effects*. For instance, to toss a coin changes its value (that can be heads or tails): this is the ontic effect. Moreover, this action also changes the mental state of any observer (this is an epistemic effect).

The hypothesis we made here is that every action can be expressed by a sequence of an *ontic action* (that is, an action having only ontic effects) and of an *epistemic action* (having effects only on the agents mental states). We only consider here a particular epistemic action: the sensing action "to observe that" ("observe" for short). To observe that p is true is only executable if p is (really) true. Note that the fact that the observe action is executable is sufficient for knowing that p is true. For instance, we can suppose that we toss a coin without looking at the coin (we close our eyes) and after that execution we observe the result (and thus, we believe that the coin is heads if it is really heads, and we believe that it is tails if it is really tails).

We say that an action is *uninformative* when its effects can be determined before its occurrence. We suppose that every ontic action is uninformative. For instance, the effect of the physical action of tossing a coin is that this coin will be heads or tails. Contrary to uninformative actions, effects of informative actions cannot be anticipated. For instance, we cannot determine the effects of the action *to test if the coin is heads or tails* (that is a sensing action) before any occurrence of this action, and even if we are aware of such an occurrence it is not sufficient to know its result. But it is not the case that every sensing action is uninformative. Recall that our action *to observe that φ is true* is executable if and only if φ is true. Thus, although it is a sensing action, it is an uninformative action (its effect can be deduced just from its occurrence). Finally, this explains why we consider in the following that every action is uninformative.

We focus here on belief change: an agent can believe that an action α has been performed whereas another action β has been really executed (the agent may misperceive the action that has really been performed). A lot of approaches were formalized with knowledge operators. As knowledge is traditionally defined as a true belief, consequences are twofold:

1. An agent has perfect knowledge about the performed action, that is:

$$K_i\langle\alpha^{-1}\rangle\top \rightarrow \langle\alpha^{-1}\rangle\top$$

(if agent i knows that the action α has just been performed then α has just been performed in the real world). For instance, if agent j tosses

a coin (noted j:*toss* in Figure 3.3), agent i knows that a tossing action has been performed by agent j.

2. Agent's knowledge cannot be erroneous with respect to the real world and thus, update operators *à la Katsuno-Mendelzon* [91][4] are sufficient to take into account changes in the world (because this agent perfectly knows the action laws). For instance, when a coin is tossed, agent i *knows* that the coin is either heads or tails (but not both together) and it cannot be wrong. (We suppose that the agent has not observed the result.) This state is represented by $\mathscr{K}_i(w')$ in Figure 3.3. As before this action he knew (for instance) that the coin was neither heads nor tails, he needs an epistemic update operator for taking into account this change in the world (represented by $\Diamond_{toss}$ leading to $\mathscr{K}_i(w')$).

3. As knowledge is necessarily coherent with the real world, no surprise can occur. Thus, sensing actions just need expansion operators *à la Alchourrón-Gärdenfors-Makinson* [5], that is, the operation for adding a new information that is consistent with the current beliefs of the agent. (More details about revision and update can be found in our PhD thesis [100].) For instance, after the execution of a tossing action, when i observes the coin (let us suppose that the coin is heads, as showed in w'' in Figure 3.3), it just need to remove the epistemic worlds where tails occurs. Thus, heads occurs in every epistemics world and, by definition, agent i knows that the coin is heads. (See $\mathscr{K}_i(w'')$ in Figure 3.3 where $+_h$ is the epistemic expansion by h.)

In our article, we use belief rather than knowledge. That is, an agent can not only have: 1) incomplete information (the world has changed but the agent does not know about that) but also 2) erroneous beliefs (about the action that has been executed in the real world) or 3) illusions (the agent believes that an action has just occurred but nothing really happens). [5] Consequently, things are more complex and we need a revision operator *à la*

[4]We recall that an update is needed when the world change (the agent need to update its beliefs in the aim to believe how is the world NOW), whereas revision is needed when the world does not change but the agent has wrong beliefs about the world.

[5]We define a set of atomic formulas $perc(\alpha, \beta)$ that is read: "the occurrence of the action α is perceived by the agent as an occurrence of the action β. We note λ the empty action (do nothing). Consequently, 1) is formalized by $perc(\alpha, \lambda)$, 2) by $perc(\alpha, \beta)$ et 3) by $perc(\lambda, \beta)$."

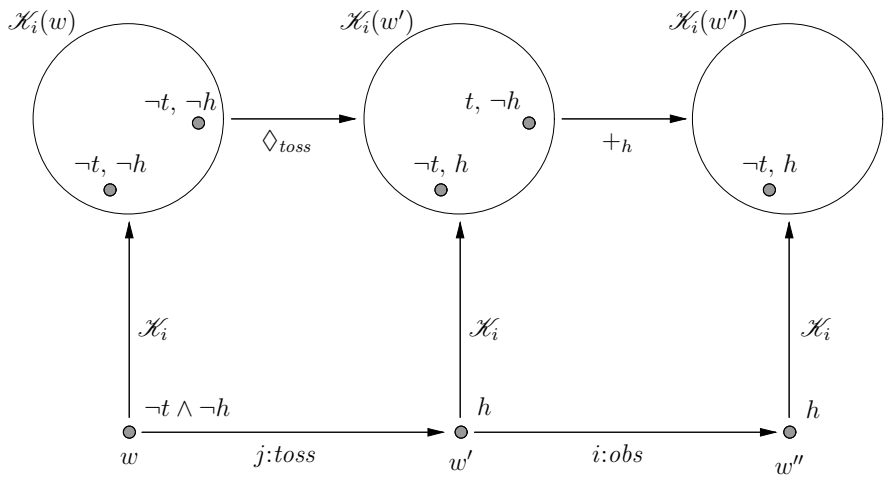

Figure 3.3: Action and knowledge using update ($\Diamond$) and expansion ($+$) operators

Alchourrón-Gärdenfors-Makinson (whereas knowledge only needs an expansion operator). For instance, let us suppose that a fixed coin (f is read "the coin is fixed") may be heads (h), or tails (t), or both together ($h \wedge t$) where the latter means that the coin has felt on its edge. Moreover, let us suppose that agent i believes the coin is not fixed. As i wrongly believes that the coin is not fixed, it does not consider that a possible state of the coin is $h \wedge t$. It follows from that wrong belief that the worlds in $\mathscr{B}_i(w')$ are not compatible with w' (see Figure 3.4). Moreover, after the observation of the coin, agent i cannot make an expansion *à la Alchourrón-Gärdenfors-Makinson*[6] because if we only keep the worlds in $\mathscr{B}_i(w')$ that are compatible with w'... we keep no world at all! Thus, agent i has no belief about nothing (or can believe everything) and it is not satisfactory. Thus, how could we built $\mathscr{B}_i(w'')$?

We partially (we only consider uninformative actions) solve this problem by enabling actions that agents believe inexecutable. The idea is to consider that, when an unexpected action is performed a mental action updates the epistemic states of this agent. From an intuitive point of view, let us consider that, when an action that we believe to be impossible happens, we cannot

[6]From a semantical point of view, it means that, when some proposition p is observed, only the epistemic worlds where p holds are kept.

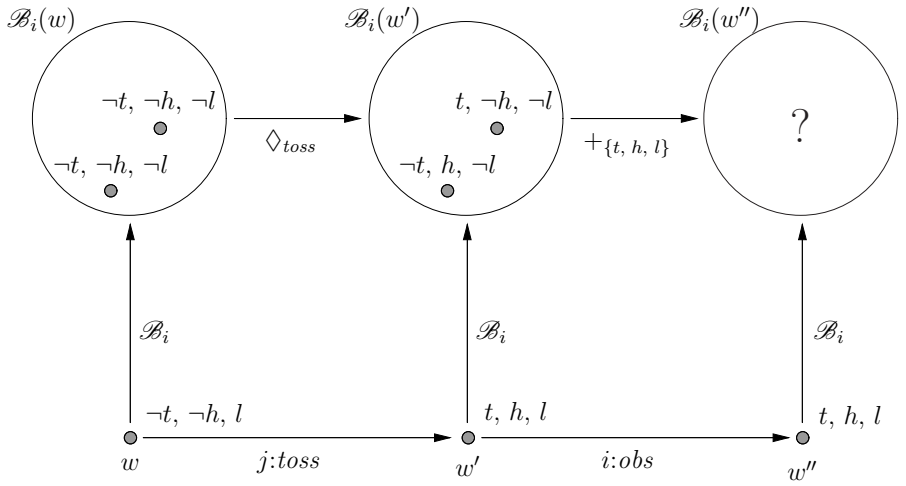

Figure 3.4: Action and belief using update ($\Diamond$) and expansion ($+$) operators: when an agent believes nothing anymore

deny the reality: we are obliged to recognize that this action has just been performed and then, that this action was possible. In this way, we update our mental states: we believe now that this action was possible.

Let us consider for instance the action *to toss a fixed coin*. The possible effects of this action are heads, tails, or heads and tails (when the coin fails on its edge). Let us suppose that agent i does not believe that the coin is fixed and from its point of view, the action of tossing a fixed coin is not executable.

We start with the fact that observation actions remain expansions *à la Alchourrón-Gärdenfors-Makinson*. Roughly speaking, the general case remains an update followed by an expansion, but when it entails that $\mathscr{B}_i(w'')$ is empty (see Figure 3.4), it means that the previous state $\mathscr{B}_i(w')$ has not been accurately built. In this case, we must apply a revision operator. This revision is a two-steps process: first, we apply on the agent's epistemic states a particular action that will enable an action that was believed not executable by the agent.

The both following axioms summarize the belief change process.

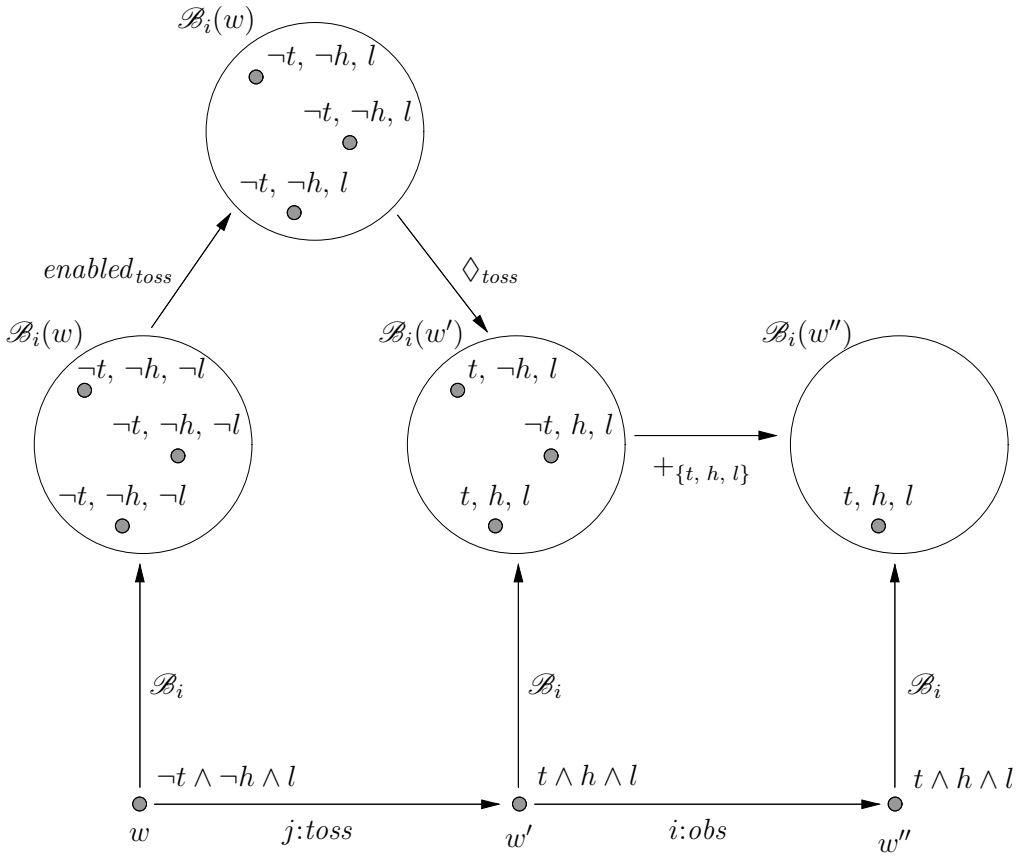

Figure 3.5: Action and belief using *enabled* operator operator

$$(perc(\alpha,\beta) \wedge \neg After_{\alpha} \bot \wedge \neg BelAfter_{\beta} \bot) \rightarrow (Feas_{\alpha}\, Bel\varphi \leftrightarrow BelAfter_{\beta}\, \varphi) \quad (\text{SSA}_1)$$

where α is an uninformative action.

$$(perc(\alpha,\beta) \wedge BelAfter_{\beta} \bot \wedge \neg After_{\alpha} \bot) \rightarrow (Feas_{\alpha}\, Bel\varphi \leftrightarrow BelAfter_{enable_{\beta}}\, After_{\beta}\, \varphi) \quad (\text{SSA}_2)$$

where α is an uninformative action.

In (SSA$_1$), the occurrence of action α is perceived as an occurrence of action β (as specified by $perc(\alpha,\beta)$), action α is executable in the real world ($\neg After_{\alpha}$) and the agent does not believe that β is inexecutable ($\neg BelAfter_{\beta} \bot$). Thus, α can be executed and the agent will believe that φ is true after its execution ($Feas_{\alpha}\, Bel\varphi$) if and only if this agent believes that after the execution of action β, φ will be true ($BelAfter_{\beta}\, \varphi$).

In (SSA$_2$) the agent still perceives an occurrence of α as being an occurrence of β ($perc(\alpha,\beta)$) but it believes now that β is inexecutable ($BelAfter_{\beta} \bot$). Under these conditions, α can be executed and the agent will believe that φ is true after its execution ($Feas_{\alpha}\, Bel\varphi$) if and only if this agent believes that after the fact that β has been enabled followed by the execution of β, φ will be true ($BelAfter_{enable_{\beta}}\, After_{\beta}\, \varphi$).

Finally, our framework allows us to define the new mental state of an agent after the execution of an action, even if this action is not correctly perceived by the agent or this agent believed that this action was not executable.

3.2 Social concepts

In the previous section we have focused on several concepts related to single cognitive agents. It has mainly concerned their mental states and both their linguistical and physical actions. As it has already been noted in Section 1.4, some criticisms have been made. Problems mainly come from the fact that the system is viewed just as a set of single agents and not as an autonomous entity having its own properties and acting by using joint actions. We need thus some concept(s) helping us to catch a set of agents as a group in its whole: we will be thus capable to deduce if some property is believed by the

group for instance or what are its goals; how this group is organized and how can this organization impact on the mental states and on the actions of each agent of the group? etc.

It is quite common to attribute some mental attitudes to groups that we call *group attitudes* in the following, such as in: "French people love good food", "Brazilian people believe that football is more important than everything else", "The government has increased taxes" or "Tom's friends offered him a very nice gift", etc. Nevertheless, as Tuomela said, to speak about a *group belief* is more a metaphor or an analogy with individual states than a mental state in the usual sense.[7]

Similarly to other concepts such as intention for instance, an interesting philosophical question is whether a given social attitude is or is not a primitive concept (that is, a concept that we could define as a complex of other concepts). For instance, when we say that "French people love good food" we want to say that a majority of individuals among French people loves good food. Thus, we could explain what we mean by "French people" with the help of French individuals and we do not need a specific group concept for describing this group attitude (reductionist definition).[8]

But thinks become more complex when we consider the well-known example of Tuomela [151]: "The Communist Party of Ruritania believes that capitalist countries will soon perish (but none of its members really believes so)". As this example requires that nobody in the group G (the Communist Party of Ruritania) individually believes the fact that p (capitalist countries will soon perish), the Tuomela's group belief should entails something like $\bigwedge_{i \in G} \neg Bel_i\, p$. But generally, reductionist definitions of group attitude operators logically entail that $\bigvee_{i \in G} Bel_i\, p$ (at least one of the members of G believes that p is true).[9] This property leads to a contradiction if we use them for formalizing the Tuomela's group belief. In other words, this example does not support a reductionist point of view of group attitudes; we need to define new concepts here.

[7] We agree with that point of view and it is the reason for which we use the *group attitude* expression rather than *group mental state*.

[8] Note that it is not necessary in this example that every French person that loves good food is aware of the fact that others French persons (do not) love good food. This is often the case for non structured groups (that is, groups having no internal organization).

[9] It is true in particular for both common belief and shared belief about p because each of these operators entails that every agent of the group believes p ($\bigwedge_{i \in G} Bel_i\, p$).

3.2.1 Groundedness as the expression of beliefs

At the origin. At the very beginning of our works on social concepts, we did want to manage interactions between agents. Two main approaches have been followed for formalizing dialogues. The *mentalist approach* (often based on a BDI logic; see [33, 122, 131] for instance) considers that a dialogue depends on the agents' mental states. It has a great predictive power but uses very strong hypotheses on the agents' internal architecture and their mental states. As said above (see Section 1.4), this approach has often been criticized (see [143, 53] for instance) because these hypotheses do not apply to open systems with heterogenous agents. The main reason is that mentalist approaches need to make hypothesis about (internal) mental states. To get around this problem, *conventional approaches* (or *social approaches*) take into account only what is public in the dialogue [36, 159, 143, 157] and describe dialogue through the notion of commitment.

But we think there is a confusion here, in the understanding of speech acts theory. In this theory, there is a clear difference between what is *expressed* by the speaker and what the speaker really *believes*. In the case of assert acts for instance, a speaker cannot assert p without *expressing* that p is true (else it would be a Moore's paradoxe[10]). It means that, if the speaker does not *express* the fact that she believes p while she is asserting p, it is not an assertion that has been performed. In other words, to express the fact that one believes p while one is asserting p is a constitutive rule of the assertive illocutionary act (see [133, Sec. 2.5] for more details about constitutive rules).

But it does not mean that the speaker really believes that p and sincerity (that is, the fact that the speaker believes p) is not required by the hearer for understanding that the speaker has asserted p. (In terms of speech act theory, the fact that the speaker believes p is neither a necessary condition nor a sufficient condition for successfully asserting p.) Thus, following speech acts theory, when the speaker asserts p, the hearer can (only) believe (at least in a first step) that the speaker has *expressed* that he believes p. But in no case the hearer believes *de facto* that p is true. (For that, we agree with

[10]The Moore's paradox is the fact that one cannot successfully assert "p is true and I do not believe p". The paradox follows from the fact that: on the one hand, the assertion entails expression of the sincerity condition about p (the speaker believes p); on the other hand, the assertion expresses the speaker believes he ignores that p. If we accept introspection then this expresses that the speaker does not believe p, and the assertion is contradictory (if we accept that beliefs are consistent).

social approaches that the hearer must make some hypotheses about the speaker, in particular he must suppose both sincerity and competence of the speaker.) But this subtle difference between what is expressed and what is believed is not made in FIPA-ACL where, if agent i asserts p then the hearer j believes p. (A slightly better variant is when the hearer believes that the speaker believes p.) This oversimplification has played a great role in the justification of commitment-based approaches that claim the (false) idea that BDI approaches of dialogue need to make hypothesis about private mental states.

The grounding concept was an attempt to show that belief and group belief are sufficient to catch what has been expressed during the dialogue and to build something like commitment stores (or expressed belief stores, here).

Definition and properties. The first article on the logic of groundedness is [57] and its extended version [56] is presented in Section 4.2.1. (It was a work with Andreas Herzig and our PhD student Benoit Gaudou.) Some principles have been generalized in [58, 62] and groundedness operators have been applied in the aim to formalize some other concepts. ([62] is presented in Section 4.2.2.) The major difference between [56] one hand and [58] or [62] for instance on the other hand, is that in [56] what is grounded does not depend on a particular group of agents (among all the agents of the system): we suppose that all the agents of the system are participating in the current dialogue.

Let AGT be the set of all the agents. We note $2^{I*} = 2^I \setminus \{\emptyset\}$ the set of all the subsets of I except the empty set, for every $I \subseteq AGT$. In [62] we say that a proposition is grounded for a group of agents if it is publicly expressed and established by all agents of the group. In the following $G_I\,\varphi$ reads "φ is grounded for the group I", where $I \in 2^{AGT*}$.[11] We note $G_{\{i\}}\,\varphi$ as $G_i\,\varphi$ that is read "φ is only grounded by agent i" or "agent i (has expressed that it) believes that φ is true".

Groundedness is an objective notion: it refers to what can be observed, and only to that. It means that this proposition φ is a part of the background of the dialogue between the agents of the group. Thus, nobody in the group can tell something that is contradictory with the propositions that are already grounded in the group, except when this agent tells explicitly that it denies

[11] In [56], G is defined as groundedness operator and $G\varphi$ is the same as $G_{AGT}\,\varphi$ in [62].

the fact that a proposition is true. In other words, the following property is valid (see [62]):

$$G_I\,\varphi \rightarrow \neg\, G_I\,\neg\varphi$$

This property is related to some rationality of the group about what is grounded. Note that φ could be of the form $G_i\,\psi$. If I represents the set of all the agents participating in a dialogue, as soon as it is grounded for the group I that agent i believes ψ, it cannot be grounded for this group that agent i does not believe that ψ is false. In other words, agent i cannot express before this group both together ψ and $\neg\psi$.

Grounding operators are close to Tuomela's view in the sense that if some proposition is grounded in a group of agents then this proposition is not necessarily believed by some agents of this group. Thus $G_I\,\varphi$ does not necessarily imply $G_{I'}\,\varphi$ for $I' \in 2^{I*}$. (In particular, it is true when I' is a singleton.) It allows to represent situations where something is grounded for the group I but not for one of its subgroups. Suppose for instance that $I = \{i, j, k\}$ and that agent i is the manager of both the agents j and k. Moreover, suppose that p reads "The company where i, j and k work has a good pay policy". Thus, we could accept a situation such that $G_I\,p$ and $G_{\{j,k\}}\,\neg p$ are true both together. (Note that groundedness does not allow to fix in which situation agents are not sincere. In other words, sincerity is not required for something be grounded.)

An interesting property concerns introspection. In [56] the following principles are valid:

$$G_I\varphi \rightarrow GG\varphi$$
$$\neg G\varphi \rightarrow G\neg G\varphi$$

It means, together with the previous property, that φ is (not) grounded iff it is grounded that φ is (not) grounded. These properties seem to fit with Tuomela's example (see page 54) because it would be very strange that something is grounded for the Party whereas it is not grounded for the Party that it is grounded for it. But in [62], a logically stronger principle holds:

$$G_I\,\varphi \rightarrow G_{I'}\,G_I\,\varphi$$
$$\neg\, G_I\,\varphi \rightarrow G_{I'}\,\neg\, G_I\,\varphi$$

where $I' \in 2^{I*}$. It means, together with the above rationality property of groundedness, that φ is (not) grounded for a group I iff it is grounded for

every subgroup I' of I that φ is (not) grounded for I. For instance, if agents i, j and k are talking to each other, it seems intuitive that, as soon as something is grounded in this group, it is grounded for each agent that it is grounded for the group. But surprisingly, it seems not to fit perfectly with Tuomela's example. Indeed, when I represents an institution, what could be I'? Members of this institution? or (some members among) its leaders? Suppose that I' are new members of the group I (the Communist Party of Ruritania) and suppose that it is grounded for the Party that p (that is, the capitalist countries will soon perish). It seems to be realistic that new members do not know every belief of the party and thus, ignore that p is grounded for I, what is contradictory with the consequent of the above properties. This problem seems to follow from the fact that something cannot be grounded without the consciousness that this thing is grounded. We will see in Section 3.2.2 that this problem does not appear with the concept of acceptance because something can be accepted by an agent (or a group of agents) both consciously and unconsciously.

Another property that can be found in [56] is $G\varphi \rightarrow GBel_i\,\varphi$ for φ factual, that is a particular case of the generalized form that can be found in [62]:

$$G_I\,\varphi \rightarrow G_I\,G_{I'}\,\varphi \qquad \text{for } \varphi \text{ objective} \tag{3.3}$$

where $I' \in 2^{I*}$. An objective formula is the same as a factual formula: it means that such a formula does not contain any modal operator.[12] This property means that, if φ is grounded for the group I, then it is grounded for I that φ is grounded for each subgroup of I (in particular, for each member of the group). In case of a dialogue, this principle is something like a normative principle: as soon as φ is grounded for the group I, nobody (or no subgroup) can claim that φ is not grounded for it. This property seems intuitive not only for groups of agents, but also for institutions where a subgroup of an institution can be viewed as a group of agents depending on this institution, or as any "sub-institution" of it (for instance, the law could be viewed as an

[12]Suppose that agent i says to the group J that j believes that p is false. Thus, as we will see in the following, $G_j\,\neg p$ has been expressed by i and $G_J\,G_i\,G_j\,\neg p$ holds. But suppose now that agent j wants to assert that it believes p. Thus, p is expressed by j and $G_J\,G_j\,p$ holds and this entails by (3.3) that $G_J\,G_i\,G_j\,p$. In other words, as soon as an agent i asserts something about the beliefs of another agent j, this agent j cannot says the converse anymore! j cannot contradicts agent i, even if the subject of the discussion is about its own private mental states! Clearly, it is unacceptable and (3.3) must be restricted.

institution, and the different courts of law could be viewed as sub-institutions of the law). In this case (or in case of big groups of agents), it could be argued that an agent (or a subgroup of agents) may ignore the fact that φ has been grounded. But note that $G_{I'}\,\neg\varphi \wedge G_I\,G_{I'}\,\varphi$ is consistent. Thus, the above property assumes that, from the point of view of a group, everything that is grounded for it is grounded for every (subset of) agent(s) of it.

Finally, we defined the following property:

$$\left(\bigwedge_{i\in I} G_I\,G_i\,\varphi\right) \rightarrow G_I\,\varphi$$

that means: if it is grounded for a group I that it is grounded for every member of this group that φ is true, then it is grounded for I that φ is true. It is an important principle that allows a common ground building. It could be viewed as a democratical consequence of a universal vote where each agent expresses its point of view. (Some variants of this principle are presented in the following.)

Let us note that groundedness is different from other objective concepts such as social commitment. To see this consider the speech act where agent i asks agent j if j can pass the salt to him. Thereafter it is established (if we assume that the speech act is well and completely understood) that i wants to know whether j is able to pass him the salt (literal meaning), or that i wants j to pass him the salt (indirect meaning). In a commitment-based approach this typically leads to a conditional commitment (or precommitment) of j to pass the salt, which becomes an unconditional commitment upon a positive reaction (see for instance [53]). In our approach we do not try to determine whether j must do such or such action or not: we just establish the facts, without any hypothesis on the agents' beliefs, goals, intentions, etc. or commitments.

Some uses of the concept of groundedness. In the rest of this section, we summarize the different contributions using the concept of groundedness. In [57] and in its final version [56], groundedness is used for formalizing Walton&Krabbe's persuasive dialogues (PPD_0). This kind of dialogue is used when participants have a conflicting belief about a given proposition. The core of this kind of dialogues is the concept of commitment: strong commitment and weak commitment. These commitments are propositional, that is, they speak about a proposition. (Thus, they are different from the *commitment in action* of Cohen&Levesque.) As these commitments are defined with

the help of our groundedness operator, each commitment is public for every member of the group, that is, for every agent taking part in the dialogue. The set of all beliefs that are grounded is called a *commitment store*.

Persuasive dialogues need specific speech acts that must be defined. For instance, assertion is used at the beginning of the dialogue for passing on a strong commitment about some proposition (this proposition will be then discussed during the dialogue). Thus, the hearer may concede this proposition or challenge it. In this last case, the agent that has assert the proposition must argue or retract its initial proposition, etc. (see [56, Section 5.1, Figure 2] for a full presentation of all the possibilities of speech acts; this article can be found page 109 of this dissertation). Both the preconditions and the effects of speech acts are formalized with the help of weak commitment or of strong commitment. Thus, effects of speech acts are public for the agents taking part in the dialogue.

But persuasive dialogue need not only particular speech acts, but also a particular protocol of communication. It means that an agent cannot perform every speech acts every time. For instance, if a proposition has just been challenged, an agent can only retract this proposition or give an argument in favor of the truth of this proposition. This protocol has been encoded into speech acts themselves. Indeed, every speech act changes the commitment store of the dialogue (that is public to every agent of the dialogue) but it can be performed only if some conditions are satisfied with respect to this commitment store. For instance, an agent can concede that p is true if and only if (both it intends to concede p and) it has not a weak commitment about p yet. These constraints are summarized in [56, Table 2] (see page 109 of this dissertation).

A complete example is given at the end of this article. [56, Table 3] gives the different states of the commitment store during the dialogue.

In [62] (see page 130 of this dissertation) we investigate the Agent Communication Language (ACL). As FIPA-ACL had been criticized in the literature it was interesting not only to formalize it with the help of our groundedness operators, but also to generalize it to groups of agents: while the addressee of a speech act is just a single person in FIPA-ACL, we distinguish the group of addresses J from the group of all bystanders K (the overhearers) such that $i \in K$ is the speaker and $J \in 2^{K\setminus\{i\}*}$. In other words, everything that has been said is grounded for the all the bystanders, whereas nothing is grounded

about mental states of some of them (those that do not speak at all). For instance, suppose there are only two agents, $K = \{i, j\}$, $J = \{j\}$ and the speaker is i. The greatest criticism addressed to FIPA concerns its *rational effect*. This effect is in fact a perlocutionary effect. For instance, the rational effect of an assertive speech act about p is that the hearer believes p. But it is clearly too strong: the fact that "an agent asserts that p" is not sufficient for the hearer believing p: maybe the hearer has strong arguments for denying p, has not heared what it has been said, believes nothing that is said by this speaker, etc. (see 19 for more details about speech acts and perlocutionary effects).

In [62] we formalize two kinds of speech act: the inform act and the request act. From speech acts theory point of view, an agent i wants to inform another agent j that φ is true when both

i. i believes that j ignores that φ is true, and

ii. i has not already informed j that φ is true (else, it would be something like a confirm act of a deny act instead).

These conditions are translated with the help of groundedness operators. Thus, when the agent i informs the group J that φ is true before the group K (that is, the speech act $\langle i, \mathsf{Inform}(J, \varphi), K\rangle$ is performed), it is required that:

i. it not grounded for K that φ is grounded for J ($\neg G_K\, G_J\, \varphi$);

ii. it is not grounded for K yet that i intends that φ is grounded for J ($\neg G_K\, \mathit{Intend}_i\, G_J\, \varphi$).

Nevertheless, speech act theory is not about dialogue and does not say anything about sequences of speech acts. In particular, suppose that agent i has asserted φ (thus, $\langle i, \mathsf{Assert}(J, \varphi), K\rangle$ has been performed), and suppose that just after that, i expresses the converse (that is, $\langle i, \mathsf{Assert}(J, \neg\varphi), K\rangle$). While these acts may be successfully performed from the point of view of speech act theory, it could be strange for J that agent i says something and its opposite just after. Thus, either agent i should justify the fact that it has said something and its converse, or i should explicitly cancel the first act before performing the second one. We have choose this last solution and then we impose the following additional precondition:

iii. it is not grounded for K yet that φ is not grounded for i ($\neg G_K \neg G_i\, \varphi$).

Thus, if agent i has already asserted $\neg\varphi$ before, then its is grounded for K that $G_i \neg\varphi$ and then $\neg G_i \varphi$ is also grounded for K, which is not consistent with the last precondition of an assertion about φ. In other words, agent i must cancel its first assertion about $\neg\varphi$ before asserting φ.

Requests of FIPA have also be rewritten. When agent i requests to agent j to do the action α, the preconditions are as follows: it is not grounded for the group K yet that j already intends to do α, and it is not grounded for K yet that i does not intend that j does α. As we generalize the speech act to a group J of addressees, we make a distinction between a request to some agents of J (request to some) and a request to every agent of J (request to each). The formal specification of these both acts can be found in [62]. We have distinguished two kinds of requests: those that are addressed to a single agent among a set of agents (such that in "May one of you close the door, please?") and those that are addressed to a group of agent for a joint action (such that in "Could you please take this table away" if we suppose that the table cannot be removed by a single agent). In [62], we give a case study using both assertions and requests in a negotiation dialogue game.

Some other results have been obtained with groundedness. For instance, commitments in action (that is, commitments *à la Cohen&Levesque*) are described in [60, 59]. Such commitments are viewed as some kinds of commitment on a proposition. This proposition represents the fact that the speaker intends to perform some action (in the case of commissive speech acts) or the fact that the speaker intends that the addressee of the (directive) speech act performs some action. Thus, in the same language we are able to represent the most useful speech acts for multi-agent systems (assertives, directives, and commissives). Of course, from a philosophical point of view, this is an oversimplified view of speech acts. For instance, we should not be able to request something from somebody if our institutional status does not allow to make such a request. (We cannot request our manager to empty the rubbish, but he can request us to do it!) Nevertheless, these speech acts have been formalized in the context of an ACL for which they are sufficient. Moreover, following [52] the different states of commitments during a dialogue (pending, canceled, violated, unset, fulfilled) have been also formalized. The added value of groundedness here, is that we are able to define both objective satisfaction and subjective satisfaction, that is, satisfaction respectively from the point of view of an agent and from the point of view of the real world

without any possible mistake.

Moreover, in [61], relations between groundedness and mutual (or common) belief have been studied. The main criticism against shared belief is that even if each agent of a group G individually believes that p is true, it does not necessarily entail that G believes p *qua* a group. Thus, we have added mutual belief in groundedness logic and we have proved that a proposition φ is grounded for a group G if and only if there is a mutual belief between the members of G about the fact that φ is grounded for a group G. The "only if" of this equivalence is justified by the remark above about shared belief. The "if" means that a group G cannot wrongly mutually believes that something is grounded for itself. Some other properties are described in [61]. This article is also a revised version of [57] and [56] about relations between group belief and choice.

3.2.2 The concept of acceptance

Acceptance *vs* groundedness. As it has been pointed in the previous section, some examples show that group belief does not fit in any case. For instance, let us suppose that φ is grounded for a group J; it is not always realistic to suppose that every agent j of J believes that φ is grounded for J (in particular when this group is very big). A concept very close to group belief is that of acceptance (see [145, 31, 152, 151] for instance). Acceptance and belief are generally differentiated by some properties [46, 69]. For instance:

- acceptance is generally considered as voluntary whereas belief is not, because acceptance involves some choice whereas we normally do not choose what we believe;
- acceptance entails that our behavior (speech acts, physical actions, etc.) is consistent with what we have accepted (for instance, if we have accepted that p is true we cannot act as if p was false) but we can act in a different way with respect to our beliefs;
- belief aims at truth but not acceptance (acceptance is guided by success or utility: when the Party of Ruritania accepts the idea that capitalism is perishing, it is because it is useful for its own ideology, it enforces the idea that their party is a good option or is better);

- it follows from that it is not needed that acceptance is shaped by evidences (whereas belief is generally defended by arguments such as perception, coherence, reasoning, etc.);
- beliefs are context-independent whereas acceptance depends on context;
- etc. (There exists some other differences and more details can be found in the above cited articles.)

Consequently, an agent can accept φ while it believes that φ is false and, conversely, it cannot accept φ whereas it believes that φ is true. Finally, let us note that acceptance may not only concern groups of agents but also individuals.

Definitions. In our works[13] acceptance is indexed by a non-empty group of agents (that can be a single agent) and by an institutional context. Even if, in our works, such a context is a single term of the object language, it is in our mind a set of formal rules (such as laws or regulations for instance) or conventions (such as habits or customs for instance). The agents that comply with these rules and conventions represent the *members* of this institution.[14] With respect to the origin of the rules, we are concerned here by social (or informal) institutions rather than legal (or formal) institutions. The former are related to the fact that no agent has a particular power allowing it to change the rules (that will apply to every agent) of the institution. For instance, friends taking part in a dialogue share both information and conventions that allow them to dialog; it is a very basic institution here (an *informal group* has been build) that relies on the support of the language institution, politeness, habits, etc. but where nobody decides the rules that apply to the group dialogue. Moreover, a more elaborate institution could be a game: the players comply with the game rules but nobody has the institutional power (that is, a power given by the institution itself, here the game) to change them (but they can accept to change the rules if nobody is again this choice). When an agent or a group of agents has such a power, we call the corresponding institution "a legal (or a formal) institution". For

[13] We mainly refer here to [63] and [110] that can be found page 135

[14] Institutions can be defined beyond any member. For instance, we can define the institution of marriage, the institution of a certificate, etc. But generally, institutions are built by agents for agents.

instance, French society is a legal institution because legislators decide which rules must be accepted by every French citizen.

Formally, $[C{:}x]\varphi$ is read "the agents in C accept that φ while functioning as group members in the institutional context x" where $C \in 2^{AGT*}$. For instance, $[C{:}\mathit{Greenpeace}]\mathit{protectEarth}$ is read "the agents in C accept that the mission of Greenpeace is to protect the Earth while functioning as activists in the context of Greenpeace" and $[i{:}\mathit{Catholic}]\mathit{PopeInfallibility}$[15] is read "the agent i accepts that the Pope is infallible while functioning as a Catholic in the context of the Catholic Church". (Examples coming from [63].)

It is important to note that such formulas do not suppose that agents are really functioning as members of given groups. In other words, the fact that $[C{:}x]\varphi$ holds does not logically imply that the agents in C are functioning as group members in the institutional context x. Thus, $[C{:}\mathit{Greenpeace}]\bot$ is read "agents in C are not functioning as group members in the institutional context of Greenpeace". Finally, $\neg[C{:}x]\bot \wedge [C{:}x]\varphi$ is read "the agent of the group C are functioning as group members in the institutional context x and they accept φ while they are functioning as group members in the context x" or shorter: "the agents in the group C accept φ *qua* group members in the institutional context x". In other words, there is some kind of conditional condition in the meaning of $[C{:}x]\varphi$ (related to the condition "while functioning as group members in the institutional context x") and this condition is fulfilled when $\neg[C{:}x]\bot$ is true. Thus, $[C{:}x]\varphi$ speaks about some kind of smooth running of agents in C while functioning as group members in the context x.

Finally, $[i{:}\lambda]\varphi$ is the same as $\mathit{Bel}_i\,\varphi$ (agent i believes that φ is true) and λ represents the private mental states context. There is no claim here that belief is some particular kind of acceptance: it is just a convenient notation. From a logical point of view, the $[i{:}\lambda]$ operators have the same properties as the Bel_i operators (KD45 modal logic). Let us note that $[C{:}\lambda]$ where C is not a singleton are not well-formed operators.

Properties. Every operator $[C{:}x]$ is a normal modal operator and thus, it satisfies the rule of necessitation and the K axiom. Moreover, these operators satisfy the following properties. The first one concerns a kind of introspective

[15] For convenience, we write $[i{:}x]$ for $[\{i\}{:}x]$.

process:

$$[C{:}x]\varphi \rightarrow [B{:}y][C{:}x]\varphi$$
$$\neg[C{:}x]\varphi \rightarrow [B{:}y]\neg[C{:}x]\varphi$$

where $B \subseteq C$ and where x and y are two institutional contexts. It means that for a given group C of agents, every subgroup $B \subseteq C$ has access about all the fact that are (not) accepted by C while functioning as group members in the institution x. Several explanations are needed here. (We focus on ne first property but similar remarks apply for the second one.)

First, two instances of the above property are $[C{:}x]\varphi \rightarrow [C{:}x][C{:}x]\varphi$ and $[i{:}\lambda]\varphi \rightarrow [i{:}\lambda][i{:}\lambda]\varphi$ where $i \in B$ (and then $i \in C$) and λ is the particular private context associated to belief. Thus, the first formula token is nothing but the introspection property for acceptance: if the agents in C accept φ while functioning as group members in the institutional context x then they accept to accept φ while functioning as group members in the institutional context x. To accept is a voluntary action and thus it would be counter-intuitive that both together $[C{:}x]\varphi$ and $\neg[C{:}x][C{:}x]\varphi$ are true. The second instance is nothing else than the well-known belief introspection (see [87]).

Second, the first property has also the following instance: $[C{:}x]\varphi \rightarrow [i{:}\lambda][C{:}x]\varphi$ where $i \in C$ and where λ is the context of private mental states. In other words, an agent i is conscious[16] of all what it accepts while functioning as a member of a group in every context. This principle is in accordance with the fact that "ignorance of the law is no excuse" (even if our framework does not give any detail about the process leading to such a belief).

Third, another instance of the above first property is $[C{:}x]\varphi \rightarrow [B{:}x][C{:}x]\varphi$ that is nothing but property $G_I\,\varphi \rightarrow G_{I'}\,G_I\,\varphi$ where $I' \subseteq I$ (see page 58) for acceptance.

Finally, as soon as $[C{:}x]\varphi$ holds, it is accepted by every subgroup B *in every context* y. For instance, suppose that $[C{:}poker]lyingIsOk$ holds, where C is a group of agents, *poker* is the context of the poker game, *lyingIsOk* is read "everybody can lie boldly". Moreover, suppose that $B \subseteq C$ and that *gs* is the context of Gentlemen Society. Thus, by the first above property we can deduce both together that $[B{:}poker][C{:}poker]lyingIsOk$ and $[B{:}gs][C{:}poker]lyingIsOk$ hold. The fact that B accepts $[C{:}poker]lyingIsOk$ while its members are functioning as poker players in the context of poker

[16]Both together with the second property and the fact that beliefs are rational (formally, we have the following property: $\neg([i{:}\lambda]\varphi \wedge [i{:}\lambda]\neg\varphi)$), we can deduce an equivalence here.

game has been justified in the previous point. But why should B also accept the fact $[C{:}poker]lyingIsOk$ while its members are functioning as gentlemen in the context of Gentlemen Society? Because there exists a kind of rationality of acceptances: if a fact is accepted in some context by a group C, it would be strange that this group or one of its subgroups B does not accept in every context that it has accepted this fact in a particular context (it would be as if this group had some kind of conflict among its choices). Let us note that neither $[C{:}x]\varphi$ nor $[B{:}y][C{:}x]\varphi$ logically entails that $[B{:}y]\varphi$. Thus, in our running example, neither $[C{:}poker]lyingIsOk$ nor $[B{:}gs][C{:}poker]lyingIsOk$ logically entails $[B{:}gs]lyingIsOk$ (that is: B accept the fact that "everybody can lie boldly" while its members are functioning as gentlemen in the context of Gentlemen Society.)

The second property of acceptance is as follows:

$$\neg[C{:}x]\bot \wedge [C{:}x]\varphi \rightarrow \neg[B{:}x]\bot \wedge [B{:}x]\varphi$$

where $B \subseteq C$. (Let us recall that x may be equal to λ if and only if C is a singleton.) This principle means that if agents in a group C accept φ *qua* group members in the institutional context x, then agents in every subgroup B of C also accept φ *qua* group members in the context x.

An instance of this property is $\neg[C{:}x]\bot \wedge [C{:}x]\varphi \rightarrow \neg[i{:}x]\bot \wedge [i{:}x]\varphi$ where $i \in C$. It means if agents in a group C accept φ *qua* group members in the institutional context x, then necessarily each agent i of this group also accepts φ *qua* member in the institutional context x. This property is related here to the fact that acceptance is (also) an individual voluntary action: acceptance from a group necessarily entails acceptance of individuals. The contraposition of this proposition $(([i{:}x]\bot \vee \neg[i{:}x]\varphi) \rightarrow ([C{:}x]\bot \vee \neg[C{:}x]\varphi))$ means that: **if** agent i (of a group C) is not functioning as an individual in the context x (that is, i finds its way about the institution x) **or** agent i does not accept a fact φ while functioning as individual in the context of x, **then** agents in C are not functioning as group members in the context x **or** the members of the group C doe not accept φ while functioning as group members in the context x. In other words,

Institutional facts and constitutive rules. In [63], we have been interested by normative and institutional facts. It has been noted that these facts are characterized at least by two features [92, 135, 150].

- Performativity: an attitude of a certain kind shared by a group of agents towards a normative or an institutional fact may contribute to the truth of a sentence describing the fact.
- Reflexivity: if a sentence describing a normative or an institutional fact is true, the relevant attitude is present.

For instance, if the agents *qua* group members accept a certain piece of paper as money (an institutional fact), then, in the appropriate context, this piece of paper is money for that group (performativity). At the same time, if it is true that a certain piece of paper is money for a group, then the agents *qua* group members accept the piece of paper as money (reflexivity).

We define an institutional fact as follows:

$$[x]\varphi \stackrel{def}{=} \bigwedge_{C \in 2^{AGT*}} [C{:}x]\varphi$$

In other words, φ is true in the institutional context x if and only if, for every group C of agents, the agents in C accept φ while functioning as group members in the institutional context x.

Moreover, φ is universally accepted if and only if φ is true in every institutional context:

$$[Univ]\varphi \stackrel{def}{=} \bigwedge_{x \in INST} [x]\varphi$$

Both $[x]$ and $[Univ]$ operator are normal modal operators (see [63] for more details).

Thus, a contextual conditional is defined as follows:

$$\varphi \stackrel{x}{\rhd} \psi \stackrel{def}{=} [x](\varphi \rightarrow \psi) \wedge \neg[Univ](\varphi \rightarrow \psi)$$

that is read "in the institutional context x, if φ then ψ". We require that $\neg[Univ](\varphi \rightarrow \psi)$ be true because the conditional must not be true in the institutional context x just because it is universally true (else, this conditional would not characteristic of x, but it would be trivially true in x). It is interesting to note that $\varphi \stackrel{x}{\rhd} \psi$ satisfies some intuitive properties of count-as conditionals as isolated in [89].

Normative facts have thus been defined as follows:

$$O(\varphi, x) \stackrel{def}{=} \neg\varphi \stackrel{x}{\rhd} \mathcal{V}$$

where $\mathcal{V}$ is a violation atom as in Anderson's reduction of deontic logic to alethic logic [8] and in dynamic deontic logic [112]. (Forbidding and permission are defined in a usual way with the help of obligation.)

Finally, we can define institutional facts and constitutive rules.[17] The former are defined by the way of a set of obligation and by a set of permissions (in the above sense). For instance, the formula $InstFact_{Italy}^{\{military\},\{vote\}}(toBeOfAge)$ stands for "being of age is an institutional fact in the context of Italy and is characterized by the permission to vote in the political elections and the obligation to fulfill the military duties".[18]

A constitutive rule of an institution is a rule that defines this institution. In other words, if the constitutive rules would be different, the institution itself would be different. (See [133, p. 50] for instance.) Constitutive rules have be defined as contextual conditionals where the consequent is an institutional fact. For instance, the formula $ConstRule_{Italy}^{\{military\},\{vote\}}(eighteen, toBeOfAge)$ stands for "being eighteen years old counts as being of age is a constitutive rule in the context of Italy and being of age is an institutional fact characterized by the permission to vote in the political elections and the obligation to fulfill the military duties".

The legislators. Following the previous works, we have introduced in [110] (see page 135) the concept of legislators. Until now, the concept of acceptance was mainly concerned by informal institutions. But in formal institutions, it is possible that a little subset of agents that are members of a given group has the power to change the institution where they evolve. We call these agents *the legislators*. Formally, $Leg(x)$ is a particular subset of all the agents that is legally responsible over institution x. These legislators function together *qua* member of the (formal) institution x, that is: $\neg[Leg(x){:}x]\bot$ holds for any institution for which $Leg(x) \neq \emptyset$.

By definition, legislators have the power to change the acceptance of all the members of their institution. In other words, it means that members of every group C accept, while functioning as group members in the context x, that if φ is accepted by every legislator in x then φ is true. Formally: for

[17] This part of our works has been published in [63] that we do not present in the following. Thus, we just give here a short conceptual summarize of the two main definitions.

[18] As noted in [63], a more precise formulation of this example needs a representation of the right relation which is, however, beyond the scope of this work. See [111] for more details.

every $C \in 2^{AGT*}$ and $x \in INST$ such that $Leg(x) \neq \emptyset$ we have the following principle:

$$[C{:}x]\left(\left(\bigwedge_{i \in Leg(x)} [i{:}x]\phi\right) \rightarrow \phi\right)$$

Note that unanimity of legislator acceptance is needed here, but we could imagine some other voting process (as just a majority of the legislators, for instance).

Finally, what is accepted by some institution must also be defined with the help of legislators:

$$[x]\varphi \stackrel{def}{=} [Leg(x){:}x]\phi$$

(Some properties and comparisons with related works can be found in [110], page 135).

Norms and social roles. In [107] (see page 183) a legislator is viewed as a particular role. Indeed, in a given institution, we can accept something while playing some role, and we can reject the same thing while playing another role. For instance, legislators in some institution x are also just members without any institutional power in this institution. Thus, acceptance operators have been generalized as follows: $[C{:}a{:}x]\varphi$ reads "the agents in group C accept that φ while playing role a together in the institution x". $[C{:}a{:}x]\bot$ is read "agents in C do not play role a together in the institution x" and $\neg[C{:}a{:}x]\bot \wedge [C{:}a{:}x]\varphi$ is read "agents in C play the role a together in institution x and they accept that φ while playing role a together in institution x" or simply "agents in C accept that φ *qua* players of role a in the institution x". It is what we call *collective acceptance* and we can define similar concepts for an agent i (*individual acceptance*). All the previous principle about acceptance have been redefined in a similar way.

As actions are often associated to roles, we have added them to the formal framework. The logic of action that has been used is the well-known dynamic logic [71]. Obligations have been here defined in a normal modal logic of type KD.

In [107], we analyse the role of legislator in details. Thus, some specific properties have been given for this role. For instance, there is only one group (of agents having a role) of legislators (see [107, Prop. (LegSum)]). This

property is related to the following: the legislators of an institution x have accepted that a proposition φ is obligatory in this institution if and only if φ is obligatory in this institution. This property is formalized as follows:

$$Leg_x Obl_x \varphi \leftrightarrow Obl_x \varphi$$

Thus, in the absence of the first property, the second could entail contradictory obligations in a same institution. Finally, let us recall that legislators are suitable only in case of formal institutions. A corollary is that in such formal institutions, there necessarily exists a group of legislators (see [107, Theorem (2b)]):

$$\bigvee_{C \in 2^{AGT*}} Leg(C,x)$$

Finally, we are now able to define norms of competence and institutional power. From the contextual conditional operator $\overset{x}{\rhd}$ (that we can define again from the legislator based definition of acceptance) we are able to define a *norm of competence*, that is, a power conferring rule $Power(a, \alpha, \varphi, x)$ that is read: in the institutional context x, the agents playing a role a (in x) have the power to ensure φ by performing the action α. For instance, the formula $Power(priest,gesture, married,church)$ is meant to stand for "in the institutional context of Catholic Church, the agents playing the role of priest (in the Church) have the power of marrying a couple by performing certain gestures". Let us note that such norm of competence is related to a role rather than to a particular agent. It means that every agent i that is accepted by the (legislators of the) institution x as both together playing the role a and performing the action α, exercises this institutional power. (See [107] for the corresponding formal definitions.) Thus, we are now able to demonstrate (see [107, Theorems 5&6]) that such exercise of institutional power can change institutions (that is, the facts that are accepted by this institution or the normative rules of this institution).

Conclusions. Our works on social concepts are also about some other themes that cannot be described here. For instance, we have defined in [109] speech acts as institutional actions where modeled institutions are social or informal institutions. In this article, we have focused on obligations and social commitments and we have provided an original reductionist characterization of these concepts, anchoring them in agents' attitudes. We have

also proposed a formal characterization of the speech act of promising: constitutive rules create the relation between an utterance (as a physical action) and the speech act of promise (as an institutional action) and allow to specify the deontic dimension of promise. Finally, relations between promise and social commitment are defined. (See also [108].)

A last point has to be mentioned. We have focused in this section in social concepts that are not reducible to a complex of individual beliefs. But, as it has already been noted above, sometimes reductionist definitions seem to better fit than non-reductionist definitions, and sometimes it is the converse. For instance, in [105] we have defined (together with Emiliano Lorini and our post-doctoral student Eunate Mayor) some concepts such as collective responsibility and collective emotions that are defined with the help of common belief (which is a reductionist group attitude). We have not given any details about these works for focussing on non reductionist concepts here.

3.3 Emotion and trust

3.3.1 Emotion

There is a large literature about emotion in philosophy [66, 144] for a long time (Plato and Aristotle have already worked on emotion), in psychology [96, 115] (more recently with William James' works), in economy [99] as well as in cognitive sciences [93]. Today, emotion is also a significant theme of computer science that concerns several areas such that: (multi-)agents systems, 3D graphical rendering, reasoning, decision making, learning, etc. Such research is justified by the increase in interactive systems together with their complexity. For offering more and more services, for giving more and more fined-grained responses, for having a more and more intuitive use, interactive systems have to understand humans more and more.

Emotion is certainly one the key concepts for reaching this aim because it is a key concept of our own behavior. Thus, one on the most important question is: why do we have emotions? Following an evolutionary point of view, if we have emotions then emotions are necessary to our survival and to our evolution. Darwin [38] has described how emotion allows humans for adapting themselves to their environment. For instance, when we have fear, then we open the mouth (and thus, we can breathe more efficiently, the

oxygen quantity that is available increases, and our body can react quickly and produce an effort immediately), our eyes are wide open (and thus, our visual acuity also increases, what is better for fast reactions), etc.. Moreover, emotion plays a phylogenetic role (snakes for instance, cause fear to most people, certainly because snakebite did cause a lot of deaths). Emotion is also concerned by regulation of social interactions. For instance, if we perform an action that harms somebody, we may feel regret and this regret will encourage us to right this wrong (see [24] for instance). Emotion shows to others our reactions before a given situation and allows them to deduce our action tendencies. Thus, emotion plays a role of social signalisation. Finally, and contrarily to instinct, emotion gives us a set of actions (the action tendencies) for reacting before a given situation where instinct just gives us a unique reaction. In other words, emotion is useful both from an individual and a social point of view.

In the following, after having presented what an emotion is, we will come back to logical and numerical models of emotions, as well as application in computer sciences. Let us note that our contributions about emotion follows from the original PhD thesis of Carole Adam, that we supervised together with Andreas Herzig and Fabrice Evrard.

What is an emotion? In his founding article [88], William James tries to define in a both rigorous and scientific manner what is an emotion. For him, emotion is the *direct consequence* of a body change resulting itself from some stimulus. More than one hundred years later, his point of view is still held up by researchers such as [146, 45, 37]. They defend the idea that some facial actions (such as frown for instance) play a central role in emotion regulation. Some experiments seem to prove that some motor expressions are able to amplify, indeed to trigger, some emotions.

But some other experiments contradict these results and tend to show that emotion triggering causes this body change. This is the thesis developed by cognitive appraisal theory [96, 115, 54, for instance] that is today the main approach of emotion through a multi-componential view (see [132] for instance). Here, emotion is viewed as an episodic phenomena having a short (but not instantaneous) duration and a dynamics. An very significant aspect of this approach is related to a subjective point of view of emotion (subjectivity of the appraisal process, differentiating and triggering process, action tendencies, etc.). The different components of emotion generally accepted in

the literature are the following:

1. *the feeling*;
2. *the psycho-physiological response* (pulse acceleration, variation in temperature, etc.);
3. *the motor expression* (of the face, the voice, the body members);
4. *action tendencies* (the different possibilities for acting);
5. *cognitive appraisal*.

The last component triggers the four first components. It is the cognitive process that triggers a *differentiated emotional response*, that is the process that defines which emotion is triggered. It entails that the other components of a triggered emotion depend on the nature of the differentiated emotional response and do not define it. (It is a point of view radically different from James's view.) The other components are thus just various physico-chemical body responses. The differentiated response is given by the (conscious or subconscious) appraisal of a given stimulus with respect to mental states (preferences, goals, moral values, beliefs, etc.). Emotion is thus viewed as a phenomenon relating to an episodic variation of (a part of) its components.

Thus, cognitive appraisal is the core of emotion. Whereas theories starting from James works distinguish basic emotions from others, we agree with some authors (see [115, pp. 25–32] for instance) which found this distinction "unacceptably vague": it is not evident for proving they are basic because at the origin of the all the others, because they appear in every culture, because they may be felt by animals, or because they are at the root of behaviors ensuring a part of our survival (evolutionary impact). For Ortony *et al.*, this distinction is founded on an illusion (see [115, Table 2.1 page 27] for more details), even if some authors keep this hypothesis in their theory. If we want to define a complexity criterion between different emotions, we should take the complexity of cognitive appraisal. Depending on the fact that appraisal process is made with respect to very primitive evolutionary instincts or on the contrary with respect to very high level concepts (such as social norms, religious values, education, traditions, conventional norms, etc.), the cognitive treatment may be more or less complex. Following this view, regret or shame will be more complex than fear for instance, because the appraisal

process is made with respect to more subtle concepts in the case of the former than in the case of the latter. In this sense, we can say that regret or shame are more complex emotions than fear.

An immediate consequence of this *appraisal-based origin* of emotion is that emotion is always about something (the object of evaluation). In this sense, we can say that the cognitive structure of emotion used during the appraisal process is a mental state that is (as belief, desire, or other mental states) Intentional (in the sense of [137]; see Section 1.2.1 for more details). If an Intentional state is related to a state or to an object of the world, what is an non-Intentional emotion? According to what we are saying, it is not an emotion. It is a state for which we cannot determine the satisfaction condition; its origin is not identified and we call such phenomena a *mood* instead of an emotion. Thus, Intentionality provides a criterion for distinguishing between an emotion (which is always about something) and a mood (that is about nothing). For instance, we can feel an emotion of disappointment related to the failure of our favorite team but we cannot be ashamed in general (this is a mood).

It is interesting to note that each component of emotion could be associated to a specific role. Thus, the feeling is in charge to inform somebody that she/he is in an emotional state. The psycho-physiological response aims to requisition all the body resources that are available. The motor expression is naturally devoted to communicate somebody's emotion to others. Finally, action tendencies play an important role in decision making by giving to individuals in a very short time what are the possible reactions that can be adopted for reacting to the present situation. Most of our contributions have been focused on the cognitive structure of emotion, that is, the structure of mental state that is necessary and sufficient for having a given emotion. Even if I have had the opportunity for working a little bit on the motor expression during the ANR project CECIL (of which I was the leader), this part of the project was nevertheless mainly allotted to the staff of Catherine Pelachaud (LTCI, Telecom ParisTech). Cognitive structure of emotion is related to appraisal process and action tendencies. Psycho-physiological response could only concern physically embodied agents. Concerning the fact that a computer will be able to (really) *feel* an emotion (in the same sense as humans feel emotions) is a hard philosophical question. Some people such as Searle for instance [138] question the fact that a machine is globally unable to reason in the sense of human reasoning. The general argument is that a computer is only a syntax machine whereas humans have a semantical

thinking. (See in particular his Chinese room example.) In other words, even if we precisely describe (with the help of different variables or operators) the feeling associated with a given emotion, a machine could not "feel" this emotion at all in the same sense that a human could feel this emotion. Nevertheless, even if this question is conceptually interesting we can also argue that maybe this lack of feeling is not necessarily: as soon as the behavior of the machine is the same as the behavior of a human, does it matter if the internal state of the machine is no the same as the internal state of the human? The remaining question seems thus to be the following: "is it possible for a machine to behave at any times in the same way as a human when this machine is a purely syntactical machine?" Thus, this is here an epistemological debate that is strongly out of the scope of the present work. This question is related to completeness of the set of the behaviors of the machine with respect the set of behaviors of human. The question is open. (Note that this question includes the intensity characterization of emotion: even if a machine can have different emotions with different intensities, does it "feel" these emotions differently?)

A question that has lead a debate in the psychological community about twenty years (see [161, 94] for the beginning of the debate and [95, 162] for its end) concerns relations between emotion and cognition. While Zajonc claims that emotion is a process separated from cognition, Lazarus claims that emotion cannot be separated from cognition. The ancient Greece and Rome (see Plato, Aristotle, Seneca for instance) had made a clear distinction between passion and reason and this distinction has been preserved until now. Zajonc, following this view, justifies his point of view be producing recent results in neurosciences showing that unconscious affective reaction can happen. Lazarus reject these results because in his view the fact that some processes are unconscious does not mean that these processes are out of the scope of cognition. His argument (with which we agree) is that if we feel an emotion then we necessarily react (consciously ou unconsciously) to a stimulus with respect to our preferences or to our morale/normative values.

In computer sciences, emotion is now a significant theme at different levels because emotion plays a fundamental role in communication. In fact, humans give continuously emotional information during an interaction. For instance just a "Hello!" together with a smile is together a very common and very economic way for informing somebody that we are happy to see him/her and that we wave him/her. We could also say: "Hello, I am happy to see you!"

but it is longer, and if we express that without any emotion expression, our sincerity will be certainly challenged. In other words, emotion information also supports our believability. Some works focus on the motor expression of emotions that concern their expression by a embodied conversational agent (ECA, see for instance [67, 116]). ECAs could also use emotions models for representing not only their own emotions but also those of the users that interact with them. (These aspects rather concern modeling of action tendencies and of cognitive appraisal.) Thus, they can take into account the users emotions in the aim of computing an adapted reaction but also for showing their own emotive states or a particular personality. The global aim is that these agents seem as believable as possible in a such manner that the user of these agent systems thinks they are interacting with human [14]. There are also some works concerned by the psycho-physiological response. Here, works more concern the analysis of physiological variations during an emotional episodic phenomenon.

Such models have been used in several implemented systems. For instance, [4] describes teaching tutorials dealing with an emotion process in the aim to increase both perseverance and commitment of students. A lot of games or ambient intelligence systems use emotion today. (In [2] we have presented with our colleagues a review of emotion in ambient intelligence.) Between the very large variety of ECAs, the well-known system EM[19] simulates the decrease of intensity emotion according to time for a given set of emotions with respect to the goals that have generated them. Another well-known emotional system is Gratch & Marsella's Affective Reasoner where agents use representations of both themselve and the other agents. Finally, GRETA [42] is a real-time 3D-agent that can express both simple and complex emotions. I am currently working with Catherine Pelachaud and her staff for allowing GRETA to express surprise and astonishment in a differentiated manner.

In the following, I focus on theoretical works on emotion. I first present logical models of emotions and then numerical models.

Logical models of emotion. From a logical point of view, we want to develop logical frames for formalizing some specific emotions, their properties, the relationship between different emotions, etc. (see for instance [3, 153]).

[19]This is a system based on Tok architecture from the Oz project. See `http://www.cs.cmu.edu/afs/cs.cmu.edu/project/oz/web/`.

The aim is to develop in a same language the description of artificial agents and their emotions. The design of such agents can moreover benefit from the fact that logic is a particularly well-adapted tool for reasoning. Thus, emotion can also be used in the reasoning and decision making processes. A logical formalization also obliges us to specify each component of emotion (such as specified in psychological models); it entails that emotion is generally logically defined with the help of complex modal operators.

Psychological definitions of emotions often characterize cognitive structures of emotion. It entails that we define emotions as a complex of mental attitudes about states of affairs or about actions. Thus, it is just a part of emotion but for convenience we use the word "emotion" for "cognitive structure of emotion". Moreover, when an emotion is formalized it is necessary for naming it. Naming emotions is always a very difficult task because an emotion is felt subjectively, in our body, in our mind, and the words we use for describing what we feel can change from one person to another. Moreover, in a same situation people can feel different emotions with different intensity. Thus, the same emotion may be named differently by two different persons and a same person can give the same name for two different emotions. In the following, the name we give to emotions are just labels representing a class or a type of emotion instead of an emotion itself. (It is also the case in Ortony *et al.* works for instance.)

A lot of works in computer sciences use the OCC theory, that is the emotion theory of Ortony *et al.* [115]. It is not necessarily the best/ideal/perfect psychological theory of emotions but it is an interesting theory from a computational point of view. A reason is that Ortony *et al.* define a small set of basic variables (desirability, goals, expectations, beliefs, moral values) and the taxonomy of emotions is entirely described by the combinatory of theses variables. Thus, it is a very attractive idea to catch a great set of emotions into a unique formal theory, especially when each variable of the original theory can be directly translated in the formal theory to a given logical operator. I have myself formalized this theory (see [3]).

With respect to the sense of "complex emotion" defined above, we can roughly define complexity of (cognitive structure of) emotions by the number of mental states used for defining them. Simple emotions can be formalized just with an epistemic attitude (such as belief for instance) and with a motivational (such as goal) or normative attitude (such as moral value). For instance, cognitive structure of joy with respect to the fact that φ is true can be defined from a belief about φ together with a goal that φ is true. Thus, if

$Bel_i\,\varphi$ reads "agent i believes that φ is true" and $Goal_i\varphi$ reads "φ is a goal of agent i", then we define the cognitive structure of i's joy about φ as:

$$Joy_i\,\varphi \stackrel{def}{=} Bel_i\,\varphi \wedge Choice_i\,\varphi$$

Following this definition, when agent i is happy that φ is true, she believes that φ while she wants that φ. For instance, Tom feels joy for having passed his exam because he believes he has passed his exam and it was his goal. In other words, Tom feels joy because the state of affairs corresponds to the state wished by him. The fact that beliefs and goals are congruent associates joy with a positive value. (We also say that joy has a positive valence.) It is not the case with sadness for instance, whose associated state of affairs is not congruent with goals. Thus, the cognitive structure of i's sadness about φ can be defined as follows:

$$Sadness_i\,\varphi \stackrel{def}{=} Bel_i\,\varphi \wedge Choice_i\,\neg\varphi \tag{3.4}$$

In other words, when agent i is sad that φ is true, she believes that φ is true while the fact that φ is false is a goal of hers. Due to the fact that both belief and goal about a state of affairs are not congruent, sadness has negative valence.

But simple emotions also concern moral values. What happens when moral values fit with beliefs and when they do not? Let $Value_i\,\varphi$ be a formula that reads "φ is a moral value of agent i".[20] Thus, when the belief about φ fits with a moral value, we name the corresponding cognitive structure "moral approbation". Otherwise, it is "moral disapprobation".

All these emotions can be summarized in Table 3.1.

Table 3.1: Simples emotions

$\wedge$	$Goal_i\varphi$	$Goal_i\neg\varphi$	$Value_i\,\varphi$	$Value_i\,\neg\varphi$
$Bel_i\,\varphi$	$Joy_i\,\varphi$	$Sadness_i\,\varphi$	$Approval_i\,\varphi$	$Disapproval_i\,\varphi$

As it is shown in [3] simple belief can be replaced by an expectation. It is a more complex kind of belief that means that probably a proposition will be true in the future while it is possible it will be false (there is a risk). Hope

[20] This is necessarily an internalized norm, as explained in Section 1.2.1.

and fear are typically built both from expectations and from goals. (See the above article for more details.) Of course, these kinds of emotion are more complex than simple emotions (see [3] for more details).

The most complex kind of emotion is certainly those based on counterfactual reasoning (see [106] for a detailed formal analysis) of the kind: "it is the case that φ whereas agent i could have prevented φ". This is a kind of responsibility that we can represent as follows:

$$Resp_i\varphi \stackrel{def}{=} \phi \wedge Chp_i\varphi$$

Here, $Chp_i\varphi$ is a modal operator of our language, but it can also be defined with the help of the STIT operator. (See [106] for more details.) There are several kind of responsibility (weak responsibility, strong responsibility, moral responsibility, etc.) and it not always obvious for determining if the kind of responsibility plays a role in the definition of a given emotion. (See [105] for a more complete analysis of responsibility.)

Thus, this kind of complex emotions are built from a belief about responsibility with respect to a goal or to a moral value (see [163] for instance). This responsibility can be attributed by an agent i to herself or to another agent. For instance, $Bel_i\, Resp_ip$ reads "agent i believes that she is responsible for p" whereas $Bel_i\, Resp_jq$ reads "agent i believes that agent j is responsible for q". These kind of beliefs can be mixed with motivational or normative mental states. For instance, if agent i believes that she is responsible for φ whereas the fact that φ is false is a goal of hers (that is: $Bel_i\, Resp_i\varphi \wedge Goal_i\neg\varphi$), she can feel regret about φ. When i believes that agent j is responsible for φ while she has φ as moral value (that is: $Bel_i\, Resp_j\varphi \wedge Value_i\,\varphi$), agent i can feel admiration about φ with respect to j. Others emotions can be defined in a similar way. (See [68] for more details. In this publication, we also speak about the expression of emotion from the point of view of speech acts theory. It was a work involving my PhD student Nadine Guiraud among others.)

I have recently worked on a new kind of complex emotion: shame. (This work began with our Master student Loïc Henry in 2011.) It is a complex emotion because it needs several beliefs, a goal, and a moral value. Following [24], agent i feels ashamed about φ before an agent j if and only if agent i believes there is a mutual belief about the following facts:

i. φ is true;

ii. if φ is true then she will be negatively evaluated with respect to a given criteria ψ;

iii. the fact that ψ should be true is shared by i and j (shared moral value);

iv. finally, agent i aims to be liked by agent j, to reflect a positive self-image.

For instance, a physician i could be ashamed about the fact she does not know a medicine before her patient j if and only if she believes there is a mutual belief about the facts: i) agent i does not know some new medicine; ii) the fact that i does not know some new medicine entails that i is not a good doctor; iii) from the point of view of both i and j, it is important to be a good doctor; iv) i aims to be liked by j (the point of view of j about i is important for i). Thus, complexity of this kind of emotion does not follow from a counter-factual reasoning, but from a causal reasoning (between φ and ψ) and from a simultaneous use of goals and moral values. (A complete analysis of shame can be found in [1].)

The importance of emotion research increases because computer sciences currently use only a small part of possibilities given by emotion, and they make that with a lot of difficulties. In implemented systems, models are often just single labels that are enabled or disabled according the needs of the system. Formal models based on logic oblige us to describe the nature of emotions in a very fined-grained manner. Thus, such analysis helps us to understand them.

3.3.2 Trust

The first time I worked on trust was at the beginning of the ANR ForTrust project. During this project, I have contributed to supervise Emilino Lorini, which was in post-doctoral studies in our LILaC group of research. Trust is used in some multi-agent systems for helping users in their decision making process while agents of these systems can be incompetent or malicious. (Users of these systems must have capacity for evaluating these agents before interacting with them.) A trust system is a system building for each user u of the system s such that an evaluation e is based on the following: results of past interactions between u and the agents of s; recommendations of some agents of s about others agents of s. The ranking of a given agent according to e can thus be interpreted as the trust that s may rationally attach to this agent.

Concerning applications, there are more and more and they are developed for a great number of users: e-business (Ebay, Amazon, etc.), big wiki (Wikipedia, Planetmath, etc.), social networks (Facebook, Twiter, etc.), web sites, shared databases, etc. By introducing trust in multi-agent systems new models have been developed having both better properties (theoretical justification of theses models) and better implemented behaviors (experimental justification). See [128] for an overview of the different trust models that have been developed.

Logical models of trust The main objective here is to formalize what is trust, what does "to trust somebody" mean, and what are mental states involved when an agent (the truster) trusts another agent (the trustee). One of the most influential models of trust is that of Castelfranchi & Falcone [25]. Contrarily to more computational approaches, their model does reduce trust to subjective probabilities computed from reputation of others: trust is defined as a kind of individual belief about some properties (capacity, intention, disposition, etc.) assigned by the truster to the trustee.

The formal analysis of trust presented in [86] follows the Castelfranchi & Falcone point of view and trust is built from four components: the truster i, the trustee j, an action of j and a goal φ of i. Thus, i trusts j for performing α in the aim to achieve φ if and only if:

- φ is a goal of i ($Goal_i\varphi$);
- i believes that j can perform α ($Bel_i\, Capable_j\,(\alpha)$);
- i believes that j, by performing α, φ will be true ($Bel_i\, After_{j:\alpha}\,\varphi$);
- i believes that j has the intention of doing α ($Bel_i\, Intend_j\,(\alpha)$).

For instance, when Mary trusts the nurse for looking after her children in the aim to be able to go to the cinema, Mary aims to be able to go to the cinema, she believes that the nurse is able to look after her children, she believes that the fact that the nurse looks after her children will fulfil her goal to be able to go to the cinema, and she also believes that the nurse intends to look after her children. In other words, trust can be formalized as follows:

$$Trust_{i,j}\,(\alpha,\varphi) \stackrel{def}{=} Goal_i\varphi \wedge Bel_i\,(Capable_j\,(\alpha) \wedge After_{j:\alpha}\,\varphi \wedge Intend_j\,(\alpha))$$

We could also define a kind of potential trust meaning that "If I need a person for performing the action α I know/believe that I can trust this person for performing α". As we are interested by relationships between emotions and trust we doe not consider this kind of trust here. (See [86] for more details.)

I have contributed myself to this definition through the ANR project ForTrust [85]. But during the PhD thesis of Manh-Hung Nguyen (which I supervised together with Jean-François Bonnefon, cognitive psychologist) this definition appeared as too weak for our needs. Indeed, as soon as we want to study relationships between trust and emotion we need to speak about trust betrayal or trust satisfaction but these concepts entail that the trustee believes that the truster trusts it (we speak about a *trust relation* rather than just *trust*). This property is not entailed by our previous works on trust during the ForTrust project. For instance, suppose that Tom aims to be presented to Lila ($Goal_{Tom} toBePresented$) while he believes that she intends to visit his neighbors ($Bel_{Tom}\, Intend_{Lila}\, visitsTomsNeighbors$) (but Lila does not know that Tom knows that she will visit his neighbors). Suppose moreover that Tom believes that Lila is capable to visit his neighbors ($Capable_{Lila}\,(visitsTomsNeighbors)$) and that after she visits his neighbors he will be presented to her ($After_{Lila:visitsTomsNeighbors}\, toBePresented$). Thus, following the above definition, Tom trusts Lila to visit his neighbors with respect to the fact that he will be presented to her. But what is the difference here between the fact that Lila will perform (without knowing it) an action serving Tom's goal, and the fact that Tom trust Lila for performing this action? We think this difference is in the "without knowing it": when the truster trusts the trustee for performance of a given action α it is necessary that the truster has more than a simple belief about the trustee's intentions of performing α; there must exist a kind of agreement between them about the fact that the trustee will perform α. Indeed, it is required that the truster believes that the trustee is reliable. But just believing that the trustee intends to perform α is not sufficient because not only can the truster have a wrong belief, but the trustee can also cancel his/her intention. Thus, it is necessary there exists a group belief between the truster and the trustee about the fact that a state where α has been performed will be true in the future.

Note it is not necessary to have a group belief about the goal of the truster. For instance, let be *possToStealCar* the proposition that is true when it is possible for John to steal Mary's car, and *leavesCar* the action of Mary to leave her car alone. Suppose now that John looks for a possibility for stealing

Mary's car ($Goal_{John} possToStealCar$) and that he believes that: Mary can leave her car alone ($Capable_{Mary}\,(leavesCar)$); after she has left her car alone it will be possible for him to steal it ($After_{Mary:leavesCar}\, possToStealCar$); Mary intends to leave her car alone (we can suppose that John see her to go to parking lot; $Intend_{Mary}\, leavesCar$). Finally, following the above definition, John trusts Mary for leaving her car alone in the aim to steal it. But if John said to Mary that he wants she leaves her car on parking lot because he wants steal it, probably Mary will not leave her car and thus, John could not trust Mary to leave her car.

This extended definition of trust is explained in the joint article [18] (see p. 252). The point is that when we trust somebody to do something, the fact that this person perform or does not perform the expected action has an emotional effect: if the trustee performed the expected action then the truster should feel an emotion of positive valence (happiness, joy or satisfaction for instance) but if the trustee does not perform the expected action, then the truster should feel an emotion of negative valence (like anger, disappointment or sadness for instance). These results are detailed in [18].

Of course, there are also numerical models of trust. We give a short overview of this area in [16].

Chapter 4

Selection of published articles

4.1 Speech acts and the dynamics of mental states

A Logic of Intention with Cooperation Principles and with Assertive Speech Acts as Communication Primitives

Andreas Herzig
IRIT
118 route de Narbonne
31062 Toulouse cedex 04, FRANCE
Andreas.Herzig@irit.fr

Dominique Longin *
IRIT
118 route de Narbonne
31062 Toulouse cedex 04, FRANCE
Dominique.Longin@irit.fr

ABSTRACT

We give a new logic of intention where (contrarily to Cohen&Levesque's approach) intention is a primitive modal operator having a non-normal possible worlds semantics. We then highlight the relation between intention and belief by a set of axioms. In our logic we formulate principles of cooperation allowing an agent to infer new intentions from his beliefs about other agents' intentions. Finally, building on results of linguistic pragmatics, we show that our cooperation principles allow to infer the effects of a yes-no question "Does A hold?" from that of an associated assertive "I have the intention to know whether A". In the same manner requests can be inferred, which form another important subclass of directives. It is the aim of this work to obtain a minimal logic that can be mechanized in a simple way.

Categories and Subject Descriptors

I.2.3 [**Deduction and Theorem Proving**]: Deduction, belief revision; I.2.4 [**Knowledge Representation Formalisms and Methods**]: Modal logic; I.2.11 [**Distributed Artificial Intelligence**]: Intelligent agents, Multiagent Systems

General Terms

Modal logic, speech act theory

Keywords

Logic of intention, belief, and action, cognitive robotics, cooperation

*Currently in Laboratoire Travail et Cognition, Maison de la Recherche, Université Toulouse-le-Mirail, 5 allées Antonio-Machado, 31058 Toulouse cedex 01.

AAMAS'02 July 15-19, 2002, Bologna, Italy

1. INTRODUCTION

In this paper, we consider a BDI architecture for agents and want to describe the dynamics of their mental states.

Starting from the 'common belief' that the frame problem has been solved [16], existing solutions have been extended in the last years to handle the dynamics of knowledge [20, 29].

Modifying these solutions in order to handle belief is not easy.[1] We have shown in [11] that the recent proposal of [28] encounters some difficulties.

Our aim being to define the simplest dynamic doxastic logic, we focus on speech acts and try to define a minimal set of speech act types. Nevertheless, agents must be able to perform assertive speech acts, queries and requests.

We use recent results in linguistic pragmatics showing how what is called *indirect speech acts* can be inferred from *literal speech acts* (i.e. the act that has been literally performed), explaining thus how e.g. the literal assertive act "I want to have the salt." expresses the speaker's request "Pass the salt!". We exploit these theoretical results to encode requests within assertive speech acts. We show that in the same way yes-no questions of the form "Does A hold?" can be simulated by the associated assertive "I have the intention to know whether A."

We need for this a well-defined notion of intention and a fine grained relation between intention and belief. We have chosen to use a *primitive* notion of intention having a non-normal modal logic (neither closed under implication nor conjunction). It is opposed to the [7, 17] approaches where intention is built from the more basic concept of goal. Our modal operator of intention has a non-normal possible worlds semantics.

We study the interaction between intention and belief and give a new axiom.

Finally, we use this framework to define some cooperation principles allowing an agent to infer new intentions from his beliefs about other agents' intentions. We analyze and formalize all the principles associated to these notions.

The paper is organized as follows: we introduce first the philosophical motivations of this work (Sect. 2). Then we present the formal framework and in particular the modal operators of belief and intention (Sect. 3), show how belief and intention can be related (Sect. 4), and sketch the possible worlds semantics for the resulting logic (Sect. 5). We give some principles of cooperation (Sect. 6). Finally we

[1] 'Knowledge' is viewed as true belief.

apply our results to examples (Sect. 7).

2. PHILOSOPHICAL MOTIVATIONS

In this section, we show how speech act theory can be used to simulate other classes of speech acts by assertive speech acts via cooperation principles.

Sometimes, by saying something, we want to say something else. Within *speech act theory* [1, 21] this idea has been exploited by Searle in [22] to construct the *theory of indirect speech acts*.

An indirect speech act is the speech act (indirectly) performed by performing another speech act (the 'direct' speech act). We also call these speech acts "literal speech act" and "non-literal speech act". Thus, the utterance of "Can you pass me the salt?" achieves the direct speech act corresponding to a yes-no question on the capacity of the hearer to pass the salt to the speaker.

Sometimes, this utterance achieves an (indirect) speech act corresponding to a request, where the speaker asks the hearer to pass him the salt.[2]

Legitimately, we may reject this dichotomy between direct and indirect speech acts. We have shown ourselves how results in psycholinguistics can attack the point of view where an indirect speech act is necessarily inferred from the direct speech act: a better point of view would be to associate different sets of effects to the same act according to the utterance context [5].

Neglecting such aspects of "cognitive realism" of the model we can nevertheless benefit from the fact that an assertive speech act allows us to extract its non-literal interpretation via cooperation principles.

It is important to note that we do not want to process indirect speech acts: we just want to exploit mechanisms underlying them in order to encode requests and yes-no queries by assertive speech acts.

In [4], Virbel extended Searle's approach of [22] by showing that indirect speech acts are performed by assertions or questions on three main types of arguments:

(a) the success conditions of the intended direct speech act (i.e. the speech act that we want to perform indirectly);

(b) reasons of doing (or not doing) the intended act;

(c) the planning of the intended act.

For example, let α be the speech act that is achieved by the utterance "Can you pass me the salt?". When that utterance is interpreted indirectly, its achievement produces an indirect speech act α' corresponding to the utterance "Pass me the salt!". The preparatory conditions of α' being that the speaker thinks that the hearer can pass him the salt, α is a yes-no question about the preparatory condition of α'.

Similarly, the sincerity condition of a request such as "Execute action β" being that the speaker wants the hearer to perform action β, an assertion of that sincerity condition ("I want you to execute β.") is a 'form of indirection' that is used to mean "Execute action β!".

Finally, a reason to ask somebody whether p is true being that one wants to know whether p is true, a request such as "Inform me whether p is true" can be performed by asserting that reason, e.g. by the utterance "I want to know if p is true".

We have thus found a way of performing requests and yes-no questions by means of an assertion. In the sequel of this paper we shall give a formal framework where such non-assertive acts can indeed be inferred from assertive acts.

[2]Note this is a possibility, not a necessity. This phenomenon has been stated as "every indirect speech act is *cancellable*" [30].

3. FORMAL FRAMEWORK

Based on the philosophical theories of Searle [23] and Bratman [2], our logic follows the tradition of Cohen & Levesque [7, 8] and Sadek [17, 18]. As the latter approaches, we aim at generalizing speech act theory into a theory of communication, and we suppose that the properties of the latter are derivable from (more general) principles of rational interaction.

Our language is of the BDI sort, containing modal operators of belief, intention, and action. It is a first-order multimodal logic with neither equality nor function symbols (although the first-order aspect is not important here), and with a possible worlds semantics in terms of accessibility relations and neighborhood functions for intention.

Atomic formulas are noted p, q, ... or $P(t_1, ..., t_n)$, and $\mathcal{ATM}$ is the set of all atomic formulas. The formulas will be denoted by A, B, ... We say that a formula is *objective* if it contains no modal operator.

***Belief.*.** Let $\mathcal{AGT} = \{i, j, k, \ldots\}$ be the set of agents. We associate a modal operator of belief Bel_i to every $i \in \mathcal{AGT}$. The formula $Bel_i A$ is read "agent i believes that A". $BelIf_i A$ is an abbreviation of $Bel_i A \vee Bel_i \neg A$, and reads "agent i knows whether A is true or not".[3] We adopt the modal logic KD45 as the logic of belief. This implies that we suppose that agents cannot entertain inconsistent beliefs, and that they are aware of their beliefs and of their disbeliefs. $Bel_{i,j} A$ is read "agents i and j mutually believe that A". Semantically, it is the reflexive transitive closure of union of dynamic accessibility relations.

***Intention.*.** Intention is a fundamental mental attitude, because it is at the origin of every voluntary action. We associate a modal operator of intention Int_i to every $i \in \mathcal{AGT}$, and read the formula $Int_i A$ as "agent i intends that A".

Intention is neither closed under logical truth, nor under logical consequence, conjunction, and material implication. We only postulate:

$$\frac{A \leftrightarrow B}{Int_i A \leftrightarrow Int_i B} \qquad (\mathrm{RE}_{Int_i})$$

This is in accordance with [2, 7, 17], but contrarily to these approaches, intention is primitive here, as in [15, 14]. In the latter two only closure under logical consequence had been given up, and we thus generalize their semantics.

We have chosen this solution for three reasons. First, building intention on top of other primitive notions such as goals or desires leads to various sophisticated notions of intention, with subtle differences between them. We have kept here only those properties of intention that are common to all of them, viz. extensionality. Second, as these definitions are rather complex, it is difficult to find complete

[3]We use the term "knows" here because "i believes if A" sounds odd.

automated theorem proving methods for them, while our analysis enables more or less standard completeness techniques and proof methods. Third and most importantly, we think that our simplified notion of intention is sufficient at least in many applications.

Speech acts.. In accordance with speech act theory, an act is represented by an illocutionary force and a propositional content [21]. As we only work with speech acts of assertive type, we do not mention the illocutionary force. Thus, an assertive speech act is a tuple of the form:

$$\langle i, j, A\rangle$$

where i is the author, j the addressee, and A is a formula representing the propositional content of the act. For example, $\langle u, s, Blue(\text{sky})\rangle$ represents the assertive speech act achieved by agent u's utterance towards agent s: "The sky is blue".

We suppose that utterances are public, i.e. although $\langle i, j, A\rangle$ is addressed to j, every other agent k perceives that act.

Action operators.. To each action α there are associated modal operators $After_\alpha$ and $Before_\alpha$. For example, $Bel_i Before_{\langle j,i,p\rangle} Bel_j p$ (the agent i believes that before informing that p, j believed that p, i.e. j is sincere w.r.t. to p).

The operators $Feasible_\alpha$ and $Done_\alpha$ are introduced by stipulating that $Feasible_\alpha A$ abbreviates $\neg After_\alpha \neg A$, and $Done_\alpha A$ abbreviates $\neg Before_\alpha \neg A$.[4]

We adopt the standard axiomatics for the fragment of dynamic logic corresponding to our language.

4. RELATIONS BETWEEN MENTAL ATTITUDES

We think that rather than the interaction between intentions and goals or desires (as studied by Cohen & Levesque and Sadek), it is the interaction between intentions and beliefs which is crucial. Most importantly, an agent must abandon his intention to achieve A as soon as he believes that A is true [7, 15, 17]. This is expressed by:

$$Int_i A \rightarrow Bel_i \neg A \qquad (\text{Rel}_{\text{IntBel}}1)$$

Combined with the (D) axiom expressing consistency of belief it entails consistency of intentions, i.e. $Int_i A \rightarrow \neg Int_i \neg A$.

Theorem 1 $Bel_i A \rightarrow \neg Int_i A$.

PROOF. By ($\text{Rel}_{\text{IntBel}}1$) and the axiom (D) of the modal logic for belief, we have $Int_i A \rightarrow \neg Bel_i A$. From this the theorem follows by contraposition. □

Theorem 2 $\neg Bel_i A \rightarrow \neg Int_i \neg Bel_i A$

PROOF. This can be shown from axiom (5) for belief together with theorem 1: an instance of that theorem is $Bel_i \neg Bel_i A \rightarrow \neg Int_i \neg Bel_i A$. Then axiom (5) being $\neg Bel_i A \rightarrow Bel_i \neg Bel_i A$, we thus obtain $\neg Bel_i A \rightarrow \neg Int_i \neg Bel_i A$. □

[4] $After_\alpha$ and $Before_\alpha$ correspond to the dynamic logic operators $[\alpha]$ and $[\alpha^{-1}]$, and $Feasible_\alpha$ and $Done_\alpha$ correspond to $\langle\alpha\rangle$ and $\langle\alpha^{-1}\rangle$.

Axiom ($\text{Rel}_{\text{IntBel}}1$) has been criticized in the literature, because it describes intention as strongly related to belief. For example, if an agent does not know if the light is off in a room, he will not be able to intend to switch it off. Formally, if p denotes "the light is off", then $\neg BelIf_i p \wedge Int_i p$ is a contradiction (by definition of $BelIf_i$ and by ($\text{Rel}_{\text{IntBel}}1$)). Generally, a "rational behavior" is to consider that the agent should go to the room, and if the light is already off, drop his intention to switch the light off (because his intention is already satisfied). Thus, we would be tempted to weaken ($\text{Rel}_{\text{IntBel}}1$) to: $Int_i A \rightarrow \neg Bel_i A$ Then an agent could not know whether p is true, and at the same time he could intend that p (e.g. see [24]). But we must keep in mind that according to Sadek [17, p. 120], intention is a mental attitude that commits us (in a persistent manner) to *achieve* a goal. Hence there are in fact two intentions here: (1) in a first step, there is the *intention to know if* the light is on or off; (2) in a second step, there is the intention to switch the light off if it is on. Generally, it might be said that is not rational to seek to achieve a goal which may already hold (although here are cases where, by caution or temporal constraints, we perform an action whose goal might already hold). Thus, generally, before intending to switch off the light, we check whether the light is on. The idea underlying ($\text{Rel}_{\text{IntBel}}1$) is that each time the agent is in doubt whether it is necessary to generate an intention (as in the previous example), he should first intend to know the state of the world. And only if this state does not satisfy this property, he will then intend to achieve it.

In the rest of this section we investigate how the interplay between $Int_i A$ and $Int_i Bel_i A$ can be formally captured.

As far as we know, the only work addressing this problem is [17], where a new mental attitude *want* is proposed (also named *potential intention*). This mental attitude abbreviates $Bel_i A \vee Int_i Bel_i A$. It follows from $Bel_i A \rightarrow (Bel_i A \vee Int_i Bel_i A)$ that if an agent believes A, then he wants A.

Instead of a want operator we here focus on the intention to believe. Here, "to intend to believe" refers to an introspective mechanism. Thus, an instance of theorem 1 is:

$$Bel_i Bel_i A \rightarrow \neg Int_i Bel_i A$$

In others words, an agent cannot want to believe A, if he believes that he already believes A.

We propose to add to ($\text{Rel}_{\text{IntBel}}1$) a second principle as formalized by the following axiom:

$$(Int_i Bel_i A \wedge Bel_i \neg A) \rightarrow Int_i A \qquad (\text{Rel}_{\text{IntBel}}2)$$

This axiom is read "if an agent believes A is false and intends to believe A then he will intend A". Suppose i intends to believe A, but actually he believes $\neg A$; then i should be prepared to act in order to change the world, which justifies, the intention that A. The other way round, if i ignores whether A is true or not, then the intention to believe that A can be held without holding the intention to act in order to bring about A. There seems to be no similar axiom in the literature. It allows us to prove that $(Int_i Bel_i A \wedge \neg Int_i A) \rightarrow \neg BelIf_i A$. Hence if i intends to believe A without intending A, then he ignores whether A.

Finally, our third constitutive property of the rational balance between intention and belief is the following axiom (that is derived in [17]):

$$Int_i A \rightarrow Int_i Bel_i A \qquad (\text{Rel}_{\text{IntBel}}3)$$

This means that an agent cannot intend A without intending to believe A. The converse should not be valid: we can intend to believe A without intending A. Suppose e.g. you ignore whether the light is off or not, and you intend to believe that it is off. In this case you are prepared to act in order to acquire that belief, typically by a sensing action (checking that it is indeed off), but you are not necessarily prepared to switch it off: the latter intention might be generated in a second stage when realizing that the light is on.

Note that it follows from ($\text{Rel}_{\text{IntBel}}1$) and ($\text{Rel}_{\text{IntBel}}3$) that ($\text{Rel}_{\text{IntBel}}2$) is an equivalence: $Int_i A \leftrightarrow Int_i Bel_i A \wedge Bel_i \neg A$.

Finally, there are two essential properties related to the agent's introspection capacity (cf. [17]):

$$Bel_i Int_i A \leftrightarrow Int_i A \qquad (\text{Rel}_{\text{IntBel}}4)$$

$$Bel_i \neg Int_i A \leftrightarrow \neg Int_i A \qquad (\text{Rel}_{\text{IntBel}}5)$$

These two axioms mean that intentions (respectively, non-intentions) of an agent are sound and complete with respect to his believed intentions (respectively, non-intentions).

Remark 1 *We have neglected here a property of intention of [7], viz. that an agent i cannot entertain the intention that A if he believes that A is always false. We have omitted this here in order not to introduce into our logic a temporal operator.*

5. SEMANTICS

As we have seen, except the operators Int_i, we only have normal modal operators. For all of our axioms characterizing them, the famous modular completeness result due to Sahlqvist [19] applies, and we get for free a possible worlds semantics for our logic based on accessibility relations.

The modal operators Int_i are non-normal. Their logic is that of a classical modal logic, having a neighborhood semantics [6]. These models can be combined with the accessibility relation models, and completeness of the resulting multi-modal logic can be proven in a fairly standard way for most of the axioms.

In [9, 10] it is shown that non-normal modal operators can be translated to normal modal logics: $Int_i A$ becomes $\neg\Box_{i,1}\neg(\Box_{i,2}A \wedge \Box_{i,3}\neg A)$, where $\Box_{i,1}$, $\Box_{i,2}$ and $\Box_{i,3}$ are normal modal operators.

We currently investigate tableau theorem proving algorithms for our logic, and we have already implemented part of the logic. In [3] the theoretical basis of the `Lotrec` generic tableau prover (which is still under development) was presented. As soon as semantical completeness is ensured, `Lotrec` offers an easy way of implementing sound and complete tableau method for our logic. The termination issue still remains to be addressed (and with it decidability).

6. COOPERATION PRINCIPLES

Generally speaking, to be cooperative w.r.t. an agent j means to contribute to the satisfaction of j's goals. While being a quite popular definition nowadays, it is nevertheless superficial. Ideally, the contribution should be balanced against a lot of other aspects, such as social rules and the cognitive capacities of the agent one is supposed to help. For example, to listen to j without interrupting him is a rule of social cooperation (one thus helps j to *better* satisfy his goals, viz. to speak), while not to answer more to his question than he can memorize is a rule of cognitive cooperation (related to the Gricean maxim of quantity [12]).

To be cooperative w.r.t. j also means to try to understand and satisfy j's ultimate goals (cf. [17] for that aspect). As we have shown in [5] accepting non-literal utterances of j can be seen as a form of cooperation, as well as adopting beliefs and intentions of j, and generating intentions with the aim of (indirectly) allowing j to satisfy his intentions.

Having in mind our aim of defining a minimal logic of a rational agent, we do not take into account here cognitive capacities (such as limited reasoning and introspection) and social rules. (Our axioms are nevertheless a priori consistent with a more refined analysis.) We thus restrict ourselves to two principles: *belief adoption* (an agent adopts the beliefs of another agent); *intention generation* (an agent generates intentions, in particular in order to aswer to the questions that have been put to him, and to correct erroneous beliefs of other agents).

If one wants to completely describe the mental state of an agent after such a belief adoption and intention generation process, one has to supplement these principles by principles of belief preservation, as studied in cognitive robotics [20, 28]. We have studied such principles before [13, 11], and do not go into the details here.

6.1 Belief adoption

When an agent i adopts a belief of another agent j he starts to believe himself what he believes j believes. Adoption must be constrained in some way in order to avoid to take over just everything another agent has uttered. We here formulate that condition in terms of *competence*: an agent adopts j's belief if he believes that j is competent at that belief. This notion has also been used by Cohen & Levesque [7] and Sadek [17]. Formally, in order to describe the competence of an agent at a formula we use a *relation of dependence* [13]: $i \overset{c}{\rightsquigarrow} p$ means that i is competent at p.

This enables us to formulate the following axiom:

$$\begin{array}{c} Bel_i A \rightarrow A \\ \text{if } i \overset{c}{\rightsquigarrow} A \text{ and } A \text{ is objective} \end{array} \qquad (\text{Adopt}_{\text{Bel}}1)$$

(Remember a formula is objective if it contains no modal operator Bel or Int.) Note that $\overset{c}{\rightsquigarrow}$ is a metalinguistic notion.

Remark 2 *As this is a logical axiom, and as our logic of belief is the logic modal KD45, by the inference rule of necessitation we obtain from ($\text{Adopt}_{\text{Bel}}1$) that an agent's competence is mutual knowledge: axiom entails that if $i \overset{c}{\rightsquigarrow} A$ then $Bel_{k_1} \ldots Bel_{k_n}(Bel_i A \rightarrow A)$ for every $\{k_1, \ldots, k_n\} \subseteq \mathcal{AGT}$.*

Another principle of belief adoption that supplements the above is the following: if j asserts A and A does not contradict k's beliefs, then k adopts A:

$$\neg Bel_k \neg A \rightarrow After_{\langle i,j,A \rangle} Bel_k A \qquad (\text{Adopt}_{\text{Bel}}2)$$

(Remember that k observes i's act because we have supposed that actions are public.) Note that k adopts A even if i is not competent at A. A similar principle has been proposed in [26].

6.2 Intention generation

Intention generation completes our principles of cooperation. Suppose $Bel_i Int_j A$. Intuitively, the difficulty is to take

into account the preceding axiom ($\text{Rel}_{\text{IntBel}}1$) in an appropriate way: i should only generate the intention to bring about A when i believes that A is currently false. We formalize this in the sequel.

When i doesn't believe that A is currently true ($\neg Bel_i A$) then i does not necessarily entertain the intention that A be true. Indeed, if moreover $\neg Bel_i \neg A$ then i doesn't know whether A is currently true or not, and it cannot be the case that $Int_i A$ because $\neg Bel_i \neg A$ implies $\neg Int_i A$. The only thing that can be guaranteed here is that i adopts *the intention to believe* that A (cf. our discussion about ($\text{Rel}_{\text{IntBel}}1$) in Sect. 4). If we rewrite this we obtain our central axiom:

$$(Bel_i Int_j A \wedge \neg Bel_i A \wedge \neg Int_i Bel_i \neg A) \rightarrow Int_i Bel_i A \qquad (\text{Gen}_{\text{Int}}1)$$

Remark 3 *As we have said, we did not include in our axioms that an agent i cannot entertain the intention that A if he believes that A is always false. This means that i must abandon his intention $Int_i A$ as soon as i starts to believe that A will always be false, and the other way round $Int_i A$ cannot be generated if i believes that A will always be false. We can constrain axiom ($\text{Gen}_{Int}1$) in order to guarantee this.*

Remark 4 *($\text{Gen}_{Int}1$) is too strong if there are more than two agents. Indeed, suppose that i cooperates with both j and k, and that i thinks j and k have contradictory intentions: $Bel_i Int_j A \wedge Bel_i Int_k \neg A$. Suppose moreover that $\neg BelIf_i A \wedge \neg Int_i \neg A \wedge \neg Int_i A$ (i.e. i doesn't bother at all about A). Then by ($\text{Gen}_{Int}1$) i generates the intentions $Int_i Bel_i A$ and $Int_i Bel_i \neg A$. But this is inconsistent according to ($\text{Rel}_{IntBel}1$).*

A way of taking into account such possible inconsistencies is to weaken ($\text{Gen}_{Int}1$) by adding to the premisses the condition that j's intention must be consistent with the intentions i attributes to the other agents:

$$(Bel_i Int_j A \wedge \neg Bel_i A \wedge C) \rightarrow Int_i Bel_i A$$

where C is a formula of the form $\neg Bel_i Int_{k_1} \neg A \wedge \ldots \wedge \neg Bel_i Int_{k_n} \neg A$. Such a condition C might also take into account priorities and preferences of i w.r.t. the intentions of his fellow agents.

Another way of weakening ($\text{Gen}_{Int}1$) is to stipulate that i cannot stay without taking position as soon as he learns that j and k have inconsistent intentions. This can be formalized by a principle such as

$$Bel_i Int_j A \rightarrow (Int_i A \vee Int_i \neg A)$$

6.3 Intention generation: derived principles

In the rest of the section we discuss two other important principles, and we show that they can be derived from our central axiom.

First of all, note that by theorem 2 the second premiss $\neg Bel_i A$ of our central axiom ($\text{Gen}_{\text{Int}}1$) ensures that i will not generate the intention to believe A if i already has a contradictory intention.

In accordance with ($\text{Rel}_{\text{IntBel}}2$) and($\text{Gen}_{\text{Int}}1$), if an agent i believes that an agent j has the intention that A be true, and i does not have the intention that A be false, then i adopts the intention that A be true. By theorem 1, if agent i believes that A is false then he cannot have the intention that A be false. Putting this together we obtain:

Theorem 3 *Axiom ($\text{Gen}_{Int}1$) implies*

$$(Bel_i Int_j A \wedge Bel_i \neg A) \rightarrow Int_i A \qquad (\text{Gen}_{\text{Int}}2)$$

Hence our first principle ($\text{Gen}_{\text{Int}}2$) says that if i believes that the world must necessarily change (in the aim to satisfy the j's intention), then j's intention is directly adopted.

PROOF. The hypothesis is $Bel_i Int_j A \wedge Bel_i \neg A$. On the one hand, $Bel_i \neg A \rightarrow \neg Bel_i A$ with axiom (D), and we thus obtain the second hypothesis of ($\text{Gen}_{\text{Int}}1$).
To establish the third hypothesis of ($\text{Gen}_{\text{Int}}1$) we proceed as follows: first, we derive $Int_i Bel_i \neg A \rightarrow Bel_i \neg Bel_i \neg A$ with ($\text{Rel}_{\text{IntBel}}1$). Then $Bel_i \neg Bel_i \neg A \rightarrow \neg Bel_i \neg A$ with axiom (5). We thus obtain $Int_i Bel_i \neg A \rightarrow \neg Bel_i \neg A$, and by contraposition $Bel_i \neg A \rightarrow \neg Int_i Bel_i \neg A$.
In consequence the hypotheses of axiom ($\text{Gen}_{\text{Int}}2$) imply those of axiom ($\text{Gen}_{\text{Int}}1$). The latter allows us to obtain $Int_i Bel_i A$:

$$(Bel_i Int_j A \wedge Bel_i \neg A) \rightarrow Int_i Bel_i A$$

Now $(Int_i Bel_i A \wedge Bel_i \neg A) \rightarrow Int_i A$ with ($\text{Rel}_{\text{IntBel}}2$), entailing ($\text{Gen}_{\text{Int}}2$). □

The second principle of intention generation stipulates that if agent i believes that agent j has the intention that A, and i believes that A is currently true, then i will generate the intention that j believe A.

Theorem 4 *Axiom ($\text{Gen}_{Int}1$) implies*

$$Bel_i Int_j A \rightarrow Int_i Bel_j A \qquad (\text{Gen}_{\text{Int}}3)$$

PROOF. ($\text{Gen}_{\text{Int}}1$) allows to derive ($\text{Gen}_{\text{Int}}2$), of which $(Bel_i Int_j Bel_j A \wedge Bel_i \neg Bel_j A) \rightarrow Int_i Bel_j A$ are an instance. As $Int_j Bel_j A$ implies $Bel_j \neg Bel_j A$ by ($\text{Rel}_{\text{IntBel}}1$), we have $Bel_i Int_j Bel_j A \rightarrow Bel_i Bel_j \neg Bel_j A$ by the principles of modal logic K. As $Bel_j \neg Bel_j A$ is equivalent to $\neg Bel_j A$ in KD45, we obtain $Bel_i Int_j Bel_j A \rightarrow Bel_i \neg Bel_j A$.

Thus we obtain $Bel_i Int_j Bel_j A \rightarrow Int_i Bel_j A$ from ($\text{Gen}_{\text{Int}}2$).

On the other hand, as $Int_j A$ implies $Int_j Bel_j A$ by ($\text{Rel}_{\text{IntBel}}3$), we have $Bel_i Int_j A \rightarrow Bel_i Int_j Bel_j A$ by the principles of modal logic K.

Finally transitivity of $\rightarrow$ allows us to conclude that $Bel_i Int_j A \rightarrow Int_i Bel_j A$. □

Remark 5 *Conditions $Bel_i Int_j A$ and $Bel_i A$ of ($\text{Gen}_{Int}3$) cannot be simultaneously true if j is competent at A. Indeed, $Bel_i Int_j A \rightarrow Bel_i Bel_j \neg A$ by ($\text{Rel}_{IntBel}1$), and if j was competent at A then we would have $Bel_i Bel_j \neg A \rightarrow Bel_i \neg A$, which cannot be the case because $Bel_i A$ and $Bel_i \neg A$ are inconsistent.*

To sum it up, our central axiom allows us to derive natural and powerful principles of cooperation. In the next section we shall show that they can be applied successfully to derive yes-no questions and requests from assertions.

7. INFERRING NON-ASSERTIVE ACTS

In this section, we illustrate by two examples how the effects of yes-no questions and request can be obtained from assertive speech acts via our cooperative principles.

7.1 The effects of assertion speech acts

In [18], Sadek describes three types of effects, that we present here in a slightly simplified version.

- The *rational effect* corresponds to the effect of the act on the addressee as expected by the speaker.
- The *intentional effect* is the speaker's intention to produce the rational effect on the addressee (this effect is related to the gricean point of view of communication).
- The *indirect effect*[5] corresponds to persistence (through the performance of the act) of the feasibility preconditions.

The *rational effect* is not directly produced by the addressee, but obtains only if the speech act's satisfaction conditions hold (see [21]). For example, if $\alpha = \langle i, j, p\rangle$ has just been performed, its rational effect is $Bel_j p$.

Note that in [18] the formalization of effects is more finedgrained than here. We aimed at good compromise between soundness of representation and its complexity, and we therefore simplified his model).

The *intentional effect* is based on the rational effect. It describes that the speaker wants the hearer to believe that the speaker intends to produce an effect. Thus, the intentional effect resulting from the performance of $\alpha = \langle i, j, p\rangle$ is: $Int_i Bel_j Int_i Bel_j p$.

Finally, the *indirect effect* is related to the preservation of the capacity precondition and of the relevance precondition. For example, the capacity precondition of $\alpha = \langle i, j, p\rangle$ is $Bel_i p$; the relevance precondition of α is $\neg Bel_i BelIf_j p$.[6]

We suppose here that action laws are part of the common beliefs. Thus, any observer of the performance of a speech act believes that the indirect effect and the intentional effect have occurred: if k observes the performance of α, then $Bel_k Int_i Bel_j\ Int_i Bel_j p \wedge Bel_k\ Bel_i p \wedge Bel_k \neg Bel_i BelIf_j p$ holds. If the addressee j (who is a particular observer of the act) comes to believe p (by inferring it from the effects of the act, his –other– beliefs and his rationality and cooperation rules), then the rational effect obtains, too. In this case the speech act is said to be satisfied.

7.2 The case of yes-no questions

Let $\alpha = \langle u, s, Int_u BelIf_u A\rangle$ be the speech act that has just been performed. Suppose that α corresponds to the utterance "u says he wants to know if A holds". As we have shown previously (Sect. 2), this speech act represents a form of indirection that can be interpreted as a yes-no question. The effects of α on the agent s are as follows (see Sect. 7.1):

1. $Bel_s Int_u Bel_s Int_u Bel_s Int_u BelIf_u A$
2. $Bel_s Bel_u Int_u BelIf_u A$
3. $Bel_s \neg Bel_u BelIf_s Int_u BelIf_u A$

These effects respectively correspond to the rational (1) and to the indirect effect (sincerity (2) and relevance (3)).

[5] "Indirect" must be understood as "the side effect of the act", and not as the effect of an indirect speech act.

[6] Sadek describes communicative acts that are not necessarily speech acts. As we treat here only speech acts, the capacity precondition can be viewed as the sincerity precondition.

If we suppose every agent is competent at his mental attitudes (as it is manifested by the axioms ($\text{Rel}_{\text{IntBel}}4$) and ($\text{Rel}_{\text{IntBel}}5$)), then

$$Bel_s Int_u BelIf_u A$$

is a consequence of (2) via ($\text{Rel}_{\text{IntBel}}4$). The principle ($\text{Rel}_{\text{IntBel}}1$) (with the standard principles of the logic KD45 for belief) entails

$$Bel_s Bel_u \neg BelIf_u A$$

By the same principles, $Bel_u \neg BelIf_u A$ is equivalent to $\neg BelIf_u A$, and then we get

$$Bel_s \neg BelIf_u A$$

Finally, giving $Bel_s Int_u BelIf_u A$ and $Bel_s \neg BelIf_u A$, ($\text{Gen}_{\text{Int}}2$) allows us to conclude $Int_s BelIf_u A$. Thus, the agent s satisfies the initial intention of the agent u, which was that s adopts the intention that u knows if A holds.

7.3 The case of requests

Let $\alpha = \langle u, s, Int_u Done_\beta \top\rangle$ be the speech act that has just been performed, where s is the author of β. Suppose it corresponds to the utterance "I want you to perform β". This speech act is an indirection that can be interpreted as a request (see Sect. 2). The effects of α on s are as follows (see Sect. 7.1):

1. $Bel_s Int_u Bel_s Int_u Bel_s Int_u Done_\beta \top$
2. $Bel_s Bel_u Int_u Done_\beta \top$
3. $Bel_s \neg Bel_u BelIf_s Int_u Done_\beta \top$

These effects respectively correspond to the intentional effect (1) and to the indirect effect ((2) and (3)).

If we suppose that every agent is competent at his own mental attitudes, then

$$Bel_s Int_u Done_\beta \top$$

is a logical consequence of (2). If we suppose that s believes that he has not just performed β, i.e.

$$Bel_s \neg Done_\beta \top$$

then he will intend to perform β (via ($\text{Gen}_{\text{Int}}2$)). To sum it up, s satisfies the initial intention of u, which was that s performs β.

Remark 6 *If β is an action that must be performed by u (and not by s), the corresponding utterance of u would be of the form "I want to perform β". We might interpret this utterance as an indirect speech act. Then u would ask s (in the allusive mode) to perform the action in his place. This would require a principle of the type "if an agent i believes that an agent j intends to perform some action, then the agent i will intend to perform this action". Thus, we can always add axioms in order to take into account more finedgrained language phenomena.*

Remark 7 *In our example, we have supposed that the agent s is aware that he had not already performed β. If we suppose now that s believes he has already performed β, he will intend*

that the agent u be aware of that (via ($Gen_{Int}3$).[7]) According to the reaction of u ("I did not [hear | understand | remember | ...]"), the agent s may perform β again (in this case, a new intention should be generated, because the first one has already been satisfied).

Finally, we could suppose s does not remember if he has already performed β. The intention generated by ($Gen_{Int}1$) should then be related to a research of s in his memory, with the aim of knowing if he has already performed β. According to the answer, he will generate an intention either via ($Gen_{Int}2$) or via ($Gen_{Int}3$).

This last case formally shows the point of view developed in Sect. 4 on the problem of switching the light in a room where we do not know if the light is on or off. In this sense, this example illustrates that the axiom ($Rel_{IntBel}1$) does not rely intention and belief in a too strong way.

8. DISCUSSION AND CONCLUSION

We have presented a minimal logic for cooperative interaction. It is based on a primitive notion of intention satisfying the principle that the intention that A implies the belief that A is currently false. We have completed the principles that have been put forward in the literature by a new one.

The only type of speech acts in our logic are assertions. We have shown how requests and yes-no questions can be inferred in this framework from particular assertions, in a way similar to the inference of indirect speech acts. Inference is via cooperation principles, the most important of which are original. We have thus shown that our minimal logic allows nevertheless to reason about communication in a cooperative environment.

Our formal framework is thus relatively simple, and facilitates completeness results and theorem proving.

In a series of papers, Shapiro et col. have added the notion of goal to the Situation Calculus. The proposals are all based on the notion of knowledge (and not belief), public actions and differ in the regression axiom for goals. As the authors themselves note, those in [26, 25] lead to so-called fanatic agents, who never abandon their goals (even when they learn that they became true). In [27] every goal A comes with a cancelling condition B associated to it. Once i has adopted A, he can abandon A when he learns that B is true. Nevertheless, other agents are still free to communicate goals with cancelling condition $\top$, which can never be abandoned.

It seems to us that the difficulties are inherent to the choice of defining the goals after an action by a successor state axiom. The latter requires expressing the resulting goals explicitly as a function of the previous mental state and the new information. This is not modular enough, in the sense that all the cognitive processes that are involved when i achieves a rational balance among his mental attitudes must be taken into account in that axiom. To witness, the three versions of the successor state axiom for goals in the different papers differ according to the underlying hypotheses concerning trust and sincerity.

9. REFERENCES

[7]Because of our semantics of the $Done_\beta$ operator, $Bel_s Done_\beta \top$ will not hold, the last act that has been performed being α (the assertive speech act of u).

[1] J. L. Austin. *How To Do Things With Words*. Oxford University Press, 1962.

[2] M. E. Bratman. *Intention, Plans, and Practical Reason*. Harvard University Press, Cambridge, MA, 1987.

[3] M. A. Castilho, L. Fariñas del Cerro, O. Gasquet, and A. Herzig. Modal tableaux with propagation rules and structural rules. *Fundamenta Informaticae*, 32(3/4):281–297, 1997.

[4] M. Champagne, R. Faure, A. Herzig, D. Longin, C. Luc, J.-L. Nespoulous, and J. Virbel. Formalisation logique de la communication non littérale à la lumière d'aperçus pragmatiques et neuropsycholinguistiques. In B. Chaib-draa and P. Enjalbert, editors, *Proc. Journées Francophones Modèles Formels de l'Interaction (MFI'01)*, volume 1, pages 31–47, 2001. 17 pages.

[5] M. Champagne and D. Longin. Non literal communication: From pragmatical to logical and psycholinguistical aspects. In *Int. Colloc. on Cognitive Sciences (ICCS'01)*, 2001. 23 pages.

[6] B. F. Chellas. *Modal Logic: an introduction*. Cambridge University Press, 1980.

[7] P. R. Cohen and H. J. Levesque. Intention is choice with commitment. *Artificial Intelligence Journal*, 42(2–3), 1990.

[8] P. R. Cohen and H. J. Levesque. Rational interaction as the basis for communication. In P. R. Cohen, J. Morgan, and M. E. Pollack, editors, *Intentions in Communication*. MIT Press, 1990.

[9] L. Fariñas del Cerro and A. Herzig. Modal deduction with applications in epistemic and temporal logic. In D. Gabbay, C. J. Hogger, and J. A. Robinson, editors, *Handbook of Logic and Artificial Intelligence*, volume 4 - Epistemic and Temporal Reasoning, pages 499–594. Oxford University Press, 1995.

[10] O. Gasquet and A. Herzig. From classical to normal modal logics. In H. Wansing, editor, *Proof Theory of Modal Logics*, number 2 in Applied Logic Series. Kluwer Academic Publishers, 1996.

[11] O. Gasquet, A. Herzig, and D. Longin. An analysis of communication in a logic of belief, intention and action. Technical Report IRIT/2001-07-R, Institut de Recherche en Informatique de Toulouse, Mar. 2001. 22 pages. Available in `www.irit.fr/ACTIVITES/LILaC/`.

[12] H. P. Grice. Logic and conversation. In J. Cole and J. Morgan, editors, *Syntaxe and Semantics: Speech acts*, volume 3 : *Speech Acts*. Academic Press, 1975.

[13] A. Herzig and D. Longin. Belief dynamics in cooperative dialogues. *Journal of Semantics*, 17(2), 2000. 20 pages.

[14] K. Konolige and M. E. Pollack. A representationalist theory of intention. In *Proc. 13th Int. Joint Conf. on Artificial Intelligence (IJCAI'93)*. Morgan Kaufmann Publishers, 1993.

[15] A. S. Rao and M. P. Georgeff. Modeling rational agents within a BDI-architecture. In J. A. Allen, R. Fikes, and E. Sandewall, editors, *Proc. Second Int. Conf. on Principles of Knowledge Representation and Reasoning (KR'91)*, pages 473–484. Morgan Kaufmann Publishers, 1991.

[16] R. Reiter. The frame problem in the Situation

Calculus: A simple solution (sometimes) and a completeness result for goal regression. In V. Lifschitz, editor, *Artificial Intelligence and Mathematical Theory of Computation: Papers in Honor of John McCarthy*, pages 359–380. Academic Press, San Diego, CA, 1991.

[17] M. D. Sadek. A study in the logic of intention. In B. Nebel, C. Rich, and W. Swartout, editors, *Proc. Third Int. Conf. on Principles of Knowledge Representation and Reasoning (KR'92)*, pages 462–473. Morgan Kaufmann Publishers, 1992.

[18] M. D. Sadek. Dialogue acts are rational plans. In M. Taylor, F. Nel, and D. Bouwhuis, editors, *The structure of mutimodal dialogue*, pages 167–188, Philadelphia/Amsterdam, 2000. John Benjamins publishing company. From ESCA/ETRW, Workshop on The Structure of Multimodal Dialogue (Venaco II), 1991.

[19] H. Sahlqvist. Completeness and correspondence in the first and second order semantics for modal logics. In S. Kanger, editor, *Proc. 3rd Scandinavian Logic Symposium*, volume 82 of *Studies in Logic*, 1975.

[20] R. Scherl and H. J. Levesque. The frame problem and knowledge producing actions. In *Proc. Nat. Conf. on AI (AAAI'93)*, pages 689–695. AAAI Press, 1993.

[21] J. R. Searle. *Speech acts: An essay in the philosophy of language*. Cambridge University Press, 1969.

[22] J. R. Searle. *Expression and Meaning*. Cambridge University Press, 1979.

[23] J. R. Searle. *Intentionality: An essay in the philosophy of mind*. Cambridge University Press, 1983.

[24] C. Seguin. *De l'action l'intention : vers une caractrisation formelle des agents*. PhD Thesis, Université Paul Sabatier, Toulouse, France, Mar. 1992.

[25] S. Shapiro and Y. Lespérance. Modeling multiagent systems with the cognitive agents specification language - a feature interaction resolution application. In C. Castelfranchi and Y. Lesprance, editors, *Intelligent Agents Vol. VII - Proc. 2000 Workshop on Agent Theories, Architectures, and Languages (ATAL-2000)*. Springer-Verlag, 2000.

[26] S. Shapiro, Y. Lespérance, and H. J. Levesque. Specifying communicative multi-agent systems with ConGolog. In *Working notes of the AAAI fall symposium on Communicative Action in Humans and Machines*, pages 75–82. AAAI Press, 1997.

[27] S. Shapiro, Y. Lespérance, and H. J. Levesque. Specifying communicative multi-agent systems. In W. Wobcke, M. Pagnucco, and C. Zhang, editors, *Agents and Multi-Agent Systems - Formalisms, Methodologies, and Applications*, pages 1–14. Springer-Verlag, LNAI 1441, 1998.

[28] S. Shapiro, M. Pagnucco, Y. Lespérance, and H. J. Levesque. Iterated belief change in the situation calculus. In *Proc. Seventh Int. Conf. on Principles of Knowledge Representation and Reasoning (KR 2000)*, pages 527–538, 2000.

[29] M. Thielscher. Representing the knowledge of a robot. In A. Cohn, F. Giunchiglia, and B. Selman, editors, *Proc. KR'00*, pages 109–120. Morgan Kaufmann, 2000.

[30] D. Vanderveken. Formal pragmatics of non literal meaning. *Linguistische Berichte*, 1997.

C&L Intention Revisited*

Andreas Herzig and **Dominique Longin**
IRIT - CNRS
118 route de Narbonne
F-31062 Toulouse cedex 04 (France)
{herzig,longin}@irit.fr

Abstract

The 1990 papers of Cohen and Levesque (C&L) on rational interaction have been most influential. Their approach is based on a logical framework integrating the concepts of belief, action, time, and choice. On top of these they define notions of achievement goal, persistent goal, and intention.

We here revisit their approach in a simplified, propositional logic, for which we give complete axiomatization.

Within that logic we study the definition of achievement goals, refining C&L's analysis. Our analysis allows us to identify the conditions under which achievement goals persist. We then discuss the C&L definition of intention as well as a variant that has been proposed by Sadek and Bretier. We argue that both are too strong and propose a weakened version.

Introduction

The fundamental role of intention in communication and more generally in interaction has been stressed by Bratman (1987; 1990). Bratman's analysis has inspired most of the authors in the literature, starting with Cohen & Levesque (1990a; 1990b) (C&L henceforth). Their approach has been taken up by Perrault (1990), Rao and Georgeff (1991; 1992), Sadek (1992), Konolige and Pollack (1993), and is the standard reference on BDI logics (Wooldridge 2000).

C&L and Sadek reduce intention to primitive concepts of *belief*, *choice*, *action*, and *time*. In contrast, intention is primitive in the other approaches. This is probably due to C&L's rather complex framework, which requires a modal predicate logic with equality and quantification over sequences of events, and includes a temporal logic with a binary 'before' operator. Moreover there is only part of the semantics: syntactical assumptions are postulated that have no semantical counterpart. Finally, the frame problem remains unsolved, and attempts to fill that gap (Perrault 1990) (Appelt & Konolige 1989) have turned out to be unsatisfactory (Herzig & Longin 2000).

In this paper we simplify and perfect C&L's approach. We first define and study a minimal propositional logic of action, time, belief, and choice (that we call ABC logic) able to support C&L's approach. We here take advantage of recent progress in reasoning about actions and beliefs and in product logics, and give a complete axiomatization. We then study the definition of achievement goals, refining the C&L analysis. Our analysis allows us to identify the conditions under which achievement goals persist. We then discuss the C&L definition of intention as well as a variant that has been proposed by Sadek. We argue that both are too strong and propose a weakened version.

The components of ABC logic are introduced in the next three sections. We then give a complete axiomatization. Within ABC logic we define achievement goals, and show under which conditions their persistence can be deduced. Finally we discuss how intentions can be defined from achievement goals.

Action and time

We here introduce a simple logic of action and time. Generally speaking, events and actions can be interpreted as transition relations on states, be it states of the world, mental states, dialogue states, or a blend of them. This is the kind of model that Dynamic Logic offers. We add to this logic a unary modal operator "henceforth".

Semantics of events and actions

We suppose there is a set of *events* $EVT = \{\alpha, \beta, \ldots\}$ and a set of *agents* $AGT = \{i, j, \ldots\}$. Actions are events that are brought about by agents. We sometimes write $i{:}\alpha$ to identify the agent of α. EVT contains *purely epistemic events* which do not change the physical world, but only the agents' mental states. Epistemic events include observations and communication actions.

The formula $[\alpha]\phi$ expresses that if α happens then ϕ holds after α. The dual $\langle\alpha\rangle\phi = \neg[\alpha]\neg\phi$ expresses that α happens and ϕ is true afterwards. Hence $[\alpha]\bot$ expresses that α does not happen, and $\langle\alpha\rangle\top$ expresses that α happens.

*Our work has benefitted from numerous discussions with colleagues, in particular with Robert Demolombe, Jérôme Lang, Philippe Balbiani, Jacques Virbel, Olivier Gasquet, Yves Lespérance, Daniel Vanderveken, Mehdi Dastani, Jan Broersen, Leon van der Torre, Joris Hulstijn. Thanks are due to Maarten Marx and Tinko Tinchev for information on the complexity of product logics, and to Hector Levesque for clarifications on the C&L approach. Part of the material in this paper has been presented at the Seventh Workshop on the Semantics and Pragmatics of Dialogue (DiaBruck 2003).

Semantics of time

To speak about sequences of more than one event we use a temporal operator $\Box$. $\Box\phi$ expresses that henceforth ϕ holds. A dual operator $\Diamond$ is defined by $\Diamond\phi = \neg\Box\neg\phi$ ('eventually ϕ').

Models have a set of possible worlds W, and a mapping

$$V : W \to (ATM \to \{0, 1\})$$

associating a valuation V_w to every $w \in W$. There are mappings

$$\mathcal{R}_\Box : W \to 2^W$$

and

$$\mathcal{R} : EVT \to (W \to 2^W)$$

associating sets of possible worlds $\mathcal{R}_\Box(w)$ and $\mathcal{R}_\alpha(w)$ every possible world w. We indentify such mappings with accessibility relations: $w\mathcal{R}_\Box w'$ iff $w' \in \mathcal{R}_\Box(w)$, etc. As usual,

$$w \models [\alpha]\phi \text{ if } w' \models \phi \text{ for every } w' \in \mathcal{R}_\alpha(w)$$

and

$$w \models \Box\phi \text{ if } w' \models \phi \text{ for every } w' \in \mathcal{R}_\Box(w)$$

With C&L we suppose:

- if $w\mathcal{R}_\alpha w'$ and $w\mathcal{R}_\beta w''$ then $w' = w''$;
- $\mathcal{R}_\Box$ is reflexive[1], transitive[2], and confluent[3];
- if $w\mathcal{R}_\alpha w'$ then $w\mathcal{R}_\Box w'$;
- if $w\mathcal{R}_\alpha w'$, $w\mathcal{R}_\Box w''$ and $w \neq w''$ then $w'\mathcal{R}_\Box w''$.

It follows from the last two conditions that events are organized in histories: if $w\mathcal{R}_\alpha w'$ and $w\mathcal{R}_\beta w''$ then $w' = w''$. From that it follows that events are deterministic. (To see this put $\beta = \alpha$.)

Our semantics is slightly weaker than C&L's. First, $\mathcal{R}_\Box$ is not necessarily linear. Second, w might be possible in the future without there being a particular sequence of actions leading to w: ϕ will be eventually true without necessarily having a sequence of actions which will achieve ϕ. This will be relevant when it comes to intentions, because an agent might believe w can be achieved without having a plan to reach w.

Mental attitudes

We now add the basic mental attitudes of belief and choice to the picture.

Semantics of belief

Under the *doxastic logics* denomination, modal logics of belief are popular in philosophy and AI, and the system KD45 is widely accepted.[4] In the models, for each agent i and possible world w there is an associated set of possible worlds $\mathcal{B}_i(w) \subseteq W$: the worlds that are compatible with i's beliefs. Hence every $\mathcal{B}_i$ is a mapping

$$\mathcal{B}_i : W \to 2^W$$

For every $i \in AGT$ there is a modal operator Bel_i, and $Bel_i\phi$ expresses that agent i believes that ϕ. The truth condition for the modal operator Bel_i stipulates that $w \models Bel_i\phi$ if ϕ holds in all worlds that are compatible with i's beliefs, i.e.

$$w \models Bel_i\phi \text{ if } v \models Bel_i\phi \text{ for every } v \in \mathcal{B}_i(w)$$

$\mathcal{B}_i$ can be seen as an accessibility relation, and it is standard to suppose that

- every relation $\mathcal{B}_i$ is serial[5], transitive, and euclidian[6].

$BelIf_i\phi$ abbreviates $Bel_i\phi \vee Bel_i\neg\phi$.

Semantics of choice

Among all the worlds in $\mathcal{B}_i(w)$ that are possible for agent i, there are some that i prefers. C&L say that i *chooses* some subset of $\mathcal{B}_i(w)$. Semantically, these worlds are identified by yet another accessibility relation

$$\mathcal{C}_i : W \to 2^W$$

$Choice_i\phi$ expresses that agent i chooses that ϕ. We sometimes also say that i *prefers* that ϕ.[7] Without surprises, $w \models Choice_i\phi$ if ϕ holds in all preferred worlds, i.e.

$$w \models Choice_i\phi \text{ if } w' \models \phi \text{ for every } w' \in \mathcal{C}_i(w)$$

We suppose that

- $\mathcal{C}_i$ is serial, transitive, and euclidian.

This differs from C&L, who only have supposed seriality, and follows Sadek's approach. The latter has argued that choice is a mental attitude which obeys to principles of introspection that correspond with transitivity and euclideanity.

Choice and belief

What is the relation between choice and belief? As said above, an agent only chooses worlds he considers possible:

- $\mathcal{C}_i(w) \subseteq \mathcal{B}_i(w)$.

Hence belief implies choice, and choice is a mental attitude that is weaker than belief. This corresponds to validity of the (Inc_{Choice_i}) principle $Bel_i\phi \to Choice_i\phi$. We moreover require that worlds chosen by i are also chosen from i's possible worlds, and vice versa:

- if $w\mathcal{B}_i w'$ then $\mathcal{C}_i(w) = \mathcal{C}_i(w')$.

(See Figure 1.)

Such a semantics validates the equivalences

$$Choice_i\phi \leftrightarrow Bel_i Choice_i\phi \quad (1)$$

$$\neg Choice_i\phi \leftrightarrow Bel_i\neg Choice_i\phi \quad (2)$$

$$Choice_i\phi \leftrightarrow Choice_i Choice_i\phi \quad (3)$$

$$\neg Choice_i\phi \leftrightarrow Choice_i\neg Choice_i\phi \quad (4)$$

The implication $Choice_i Bel_i\phi \to Choice_i\phi$ is also valid, but not the converse.

[1] For every $w \in W$, $w\mathcal{R}_\Box w$.

[2] If $w_1\mathcal{R}_\Box w_2\mathcal{R}_\Box w_3$ then $w_1\mathcal{R}_\Box w_3$.

[3] If $w\mathcal{R}_\Box w_1$ and $w\mathcal{R}_\Box w_2$ then there is a w_3 such that $w_1\mathcal{R}_\Box w_3$ and $w_2\mathcal{R}_\Box w_3$.

[4] The most important criticism that has been made to KD45 is that it accepts omniscience, i.e. an agent's beliefs are closed under tautologies, conjunction, and logical consequences. In particular the latter point, viz. that an agent believes all the consequences of his beliefs, has been considered to be unrealistic. We here accept omniscience to simplify the framework.

[5] For every $w \in W$, $\mathcal{B}_i \neq \emptyset$

[6] for all $w \in W$, if $v, v' \in \mathcal{B}_i(w)$ then $v' \in \mathcal{B}_i(v)$ and $v \in \mathcal{B}_i(v')$.

[7] While C&L use a modal operator 'goal' (probably in order to have a uniform denomination w.r.t. the different versions of goals they study), it seems more appropriate to us to use the term 'choice'.

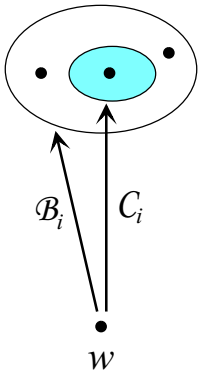

Figure 1: Belief and choice

The kinematics of mental attitudes

Several proposals were made in the beginning of the 90s concerning the relation between action and belief. They built on what was state of the art in the reasoning-about-actions field in the 80s, and used complex default or autoepistemic logics (Perrault 1990; Appelt & Konolige 1989). In the beginning of the 90s, Scherl and Levesque (1993) have proposed simple principles that can be integrated easily into the original C&L framework, which is what we undertake here.

We first make some hypotheses on the perception of events. Then we state general principles governing relationships between belief, choice, action and time.

Hypotheses on perception

We suppose that an event occurs iff every agent i perceives it. More precisely, we suppose that i's perception is correct (in the sense that if i believes that α has occurred then α indeed occurred) and complete (in the sense that if α occurs then α is perceived by i). Hence event occurrences are public.

HYPOTHESIS. *All event occurrences are perceived correctly and completely by every agent.*

We note that this hypothesis just aims at simplifying our exposition, and that misperception can be integrated following ideas of Bacchus et al. (1995; 1999) and Baltag et col. (1998; 2000).

While an agent perceives the occurrence of an event, or more precisely of an event token, we suppose that he does not learn anything beyond that about the event's particular effects. We therefore define *uninformative events* as event tokens whose outcome is not perceived by the agents. When an agent learns that such an event has occurred, he is nevertheless able to predict its results according to the action laws he believes to hold. Consider e.g. the action of tossing a coin. Suppose the agent learns that toss has occurred. As he cannot observe the effects, he predicts them in an *a priori* way, according to his mental state and the action laws. The agent might thus be said to 'mentally execute' toss. After toss he believes that $Heads \vee Tails$ holds, but neither believes $Heads$ nor $Tails$. It is only the observation that the coin fell heads which may make the agent start to believe that $Heads$.

We suppose the observation of ϕ never occurs when ϕ is false. To learn that the observation of ϕ has occurred means to learn that ϕ (supposing observations are reliable). Thus, observation actions are uninformative: all the relevant information is encoded in the notification of the event occurrence. Then to take into account the observation of ϕ amounts to incorporate ϕ into $\mathcal{B}_i(w)$.

In the same way, we can suppose that i's action of informing that ϕ is uninformative (both for the speaker i and the hearer). There are perception actions which do not satisfy our hypothesis, such as *testing-if-*ϕ. Such tests can nevertheless be reduced to uninformative actions: *testing-if-*ϕ is the nondeterministic composition of *observing-that-*ϕ and *observing-that-*$\neg\phi$.

HYPOTHESIS. *All events are uninformative.*

Our second hypothesis is deeper than the first: without presenting a formal proof here, we suppose that every event can be constructed from uninformative events by means of dynamic logic nondeterministic composition "$\cup$" and sequencing ";". For example the everyday action of tossing corresponds to the complex $\mathsf{toss}; (\mathsf{observe}\, Heads \cup \mathsf{observe}\, Tails)$. In fact such a hypothesis is often made in reasoning about actions, e.g. in (Scherl & Levesque 1993) or (Shapiro *et al.* 2000, footnote 10).

Mental attitudes and action

Suppose the actual world is w, and some event α occurs leading to a new actual world w'. Which worlds are possible for agent i at w'? According to Moore (1985) and Scherl and Levesque (1993; 2003), i makes 'mentally happen' α in all his worlds $v \in \mathcal{B}_i(w)$, and then collects the resulting worlds $\mathcal{R}_\alpha(v)$ to form the new belief state. We thus have $\mathcal{B}_i(w') = (\mathcal{R}_\alpha \circ \mathcal{B}_i)(w) = \bigcup_{v \in \mathcal{B}_i(w)} \mathcal{R}_\alpha(v)$. This identity must be restricted in order to keep i's beliefs consistent, i.e. to avoid $\mathcal{B}_i(w') = \emptyset$. We thus obtain:

- If $w \mathcal{R}_\alpha w'$ and $(\mathcal{R}_\alpha \circ \mathcal{B}_i)(w) \neq \emptyset$
  then $\mathcal{B}_i(w') = (\mathcal{R}_\alpha \circ \mathcal{B}_i)(w)$.

This relies on our hypothesis that events are uninformative: apart from the mere occurrence of α agent i should learn nothing about α's particular effects that obtain in w', and $\mathcal{B}_i(w')$ only depends on $\mathcal{B}_i(w)$ and α.

Note that such an explanation is in accordance with our hypotheses. Syntactically, this makes the principle of no forgetting (NF_{Bel_i}) $Bel_i[\alpha]\phi \wedge \neg Bel_i[\alpha]\bot \rightarrow [\alpha]Bel_i\phi$ valid, as well as the dual principle of no learning (NL_{Bel_i}) $[\alpha]Bel_i\phi \wedge \neg[\alpha]\bot \rightarrow Bel_i[\alpha]\phi$.

How do an agent's choices evolve? We recall that for each possible world there is an associated temporal structure (its history). Therefore agent i's choices concern not only possible states of the world, but also possible histories. We therefore suppose that i's preferences after α are just the images

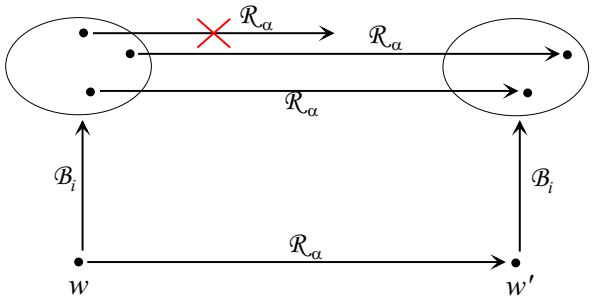

Figure 2: Action and belief

by α of its preferred worlds before α. Just as for belief, this identity must be restricted in order to keep i's choices consistent. We thus obtain the constraint:

- If $w\mathcal{R}_\alpha w'$ and $(\mathcal{R}_\alpha \circ \mathcal{C}_i)(w) \neq \emptyset$ then $\mathcal{C}_i(w') = (\mathcal{R}_\alpha \circ \mathcal{C}_i)(w)$.

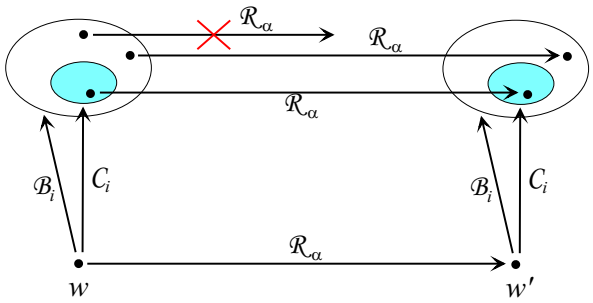

Figure 3: Action, belief, and choice

Again, note that such an explanation is in accordance with our hypotheses. Syntactically, this makes valid the principle (NF$_{Choice_i}$) $Choice_i[\alpha]\phi \wedge \neg Choice_i[\alpha]\bot \rightarrow [\alpha]Choice_i\phi$, and (NL$_{Choice_i}$) $[\alpha]Choice_i\phi \wedge \neg[\alpha]\bot \rightarrow Choice_i[\alpha]\phi$.

Mental attitudes and time

Which constraints can be formulated on Bel_i and $\Box$?

First, note that from (NF$_{Bel_i}$) it follows that $Bel_i\Box\phi \wedge \neg Bel_i[\alpha]\bot \rightarrow [\alpha]Bel_i\Box\phi$, i.e. beliefs about invariants persist as long as there are no surprises.

What about a 'no forgetting' principle for the temporal operator $Bel_i\Box\phi \rightarrow \Box Bel_i\phi$? In fact this would be too strong: suppose that for some reason, i wrongly believes that some object is broken and cannot be repaired. We thus have $Bel_i\Box\neg Broken$, which together with such a principle would imply $\Box Bel_i\neg Broken$. Which is absurd: imagine e.g. i learns that the object is in fact not broken. Then such a no forgetting principle would forbid any belief revision.

Only weaker identities can be motivated here: for each of i's possible worlds v, if u' is possible for i in some world u in the future of v then there is a world v' possible for i such that u' is in its future. And vice versa:

- if $w\mathcal{B}_i v$ then $(\mathcal{R}_\Box \circ \mathcal{B}_i)(v) = (\mathcal{B}_i \circ \mathcal{R}_\Box)(v)$

This constraint can also be interpreted as a form of introspection through time. Indeed, the introspection principles for belief correspond to $\mathcal{B}_i \circ \mathcal{B}_i = \mathcal{B}_i$, and it can be shown that due to transitivity and euclideanity of $\mathcal{B}_i$ our condition is equivalent to $\mathcal{B}_i \circ \mathcal{R}_\Box \circ \mathcal{B}_i = \mathcal{B}_i \circ \mathcal{R}_\Box$. Note that corresponding principles of negative introspection cannot be motivated.

Similar to belief we impose for choice:

- if $w\mathcal{C}_i v$ then $(\mathcal{R}_\Box \circ \mathcal{C}_i)(v) = (\mathcal{C}_i \circ \mathcal{R}_\Box)(v)$

This makes the principle (Inv$_{Choice_i}$) $Choice_i(\Box Choice_i\phi \leftrightarrow Choice_i\Box\phi)$ valid. It follows that $Choice_i\Box Choice_i\phi \leftrightarrow Choice_i\Box\phi$, which says that if an agent prefers ϕ to be invariant then he chooses that he will always prefer ϕ, and vice versa.

Comments: revision of beliefs and choices

Our conditions say nothing about i's beliefs after a surprising action occurrence, i.e. when $(\mathcal{R}_\alpha \circ \mathcal{B}_i)(w) = \emptyset$. In this case i must revise his beliefs. Integrations of belief revision into a logic of action and belief have been proposed in (Shapiro *et al.* 2000). In (Herzig & Longin 2002) we have proposed an alternative based on updating by the preconditions of α. It amounts to suppose that our language contains not only modal action operators $[\alpha]$, but also *update operators* $[upd(\phi)]$, for every formula ϕ. In the original paper such operations were seen as particular actions. Here we have to separate them because our semantics is in terms of histories, and at most one action happens at a given w, while we would like to allow several updates leaving w.

Our conditions do not constrain either i's choices when $(\mathcal{R}_\alpha \circ \mathcal{C}_i)(w) = \emptyset$, i.e. after an unwanted action occurrence. Then i has to revise his choices.

There are two cases. First, if $Choice_i[\alpha]\bot$ and $Bel_i[\alpha]\bot$ then a surprising event has occurred, and the agent has to revise both his beliefs and his choices. We think that in this case our account of belief revision in (Herzig & Longin 2002) can be extended to choice revision. In the second case we have $Choice_i[\alpha]\bot$ and $\neg Bel_i[\alpha]\bot$. Then i did not believe the event was impossible, but preferred so. Devices such as a preference relation have to be integrated here, and we leave a more detailed investigation to future work.

Completeness theorem

We have defined the semantics of a basic logic of action, belief, and choice. To sum it up, our models have the form $\langle W, \mathcal{B}, \mathcal{C}, \mathcal{R}, \mathcal{R}_\Box, V\rangle$, where W is a set of possible worlds, $\mathcal{B}$ and $\mathcal{C}$ associate accessibility relations to every agent, $\mathcal{R}$ associates an accessibility relation to every action, $\mathcal{R}_\Box$ is the accessibility relation for $\Box$, and V associates a valuation to every possible world. We call ABC *models* the set of models satisfying all the constraints imposed in the three preceding sections, and write $\models_{ABC} \phi$ if ϕ is valid in ABC models. We write $\mathcal{S} \models_{ABC} \phi$ if ϕ is a logical consequence of the set of formulas $\mathcal{S}$ in ABC models.

We give now an axiomatization of ABC. We suppose the axioms and inference rules of the basic normal modal logic

K for every modal operator,[8] plus the following:

$\neg(Bel_i\phi \wedge Bel_i\neg\phi)$ (D_{Bel_i})
$Bel_i\phi \rightarrow Bel_iBel_i\phi$ ($\mathsf{4}_{Bel_i}$)
$\neg Bel_i\phi \rightarrow Bel_i\neg Bel_i\phi$ ($\mathsf{5}_{Bel_i}$)
$\neg(Choice_i\phi \wedge Choice_i\neg\phi)$ (D_{Choice_i})
$Choice_i\phi \rightarrow Bel_iChoice_i\phi$ (PI_{Choice_i})
$\neg Choice_i\phi \rightarrow Bel_i\neg Choice_i\phi$ (NI_{Choice_i})
$Bel_i\phi \rightarrow Choice_i\phi$ (Inc_{Choice_i})
$\Box\phi \rightarrow \phi$ ($\mathsf{T}\Box$)
$\Box\phi \rightarrow \Box\Box\phi$ ($\mathsf{4}\Box$)
$\Diamond\Box\phi \rightarrow \Box\Diamond\phi$ ($\mathsf{Confl}\Box$)
$\Box\phi \rightarrow [\alpha]\phi$ ($\mathsf{Inc}_{[\alpha]}$)
$\langle\alpha\rangle\phi \rightarrow [\beta]\phi$ (Hist_1)
$\Diamond\phi \rightarrow (\phi \vee [\alpha]\Diamond\phi)$ (Hist_2)
$Bel_i[\alpha]\phi \wedge \neg Bel_i[\alpha]\bot \rightarrow [\alpha]Bel_i\phi$ (NF_{Bel_i})
$[\alpha]Bel_i\phi \wedge \neg[\alpha]\bot \rightarrow Bel_i[\alpha]\phi$ (NL_{Bel_i})
$Choice_i[\alpha]\phi \wedge \neg Choice_i[\alpha]\bot \rightarrow [\alpha]Choice_i\phi$ (NF_{Choice_i})
$[\alpha]Choice_i\phi \wedge \neg[\alpha]\bot \rightarrow Choice_i[\alpha]\phi$ (NL_{Choice_i})
$Bel_i(\Box Bel_i\phi \leftrightarrow Bel_i\Box\phi)$ (Inv_{Bel_i})
$Choice_i(\Box Choice_i\phi \leftrightarrow Choice_i\Box\phi)$ (Inv_{Choice_i})

Some comments are in order.

(PI_{Choice_i}) is an axiom of positive introspection for choice similar to ($\mathsf{4}_{Bel_i}$) and (NI_{Choice_i}) is the negative version.

Axiom (Hist_1) implies determinism of every α: $\langle\alpha\rangle\phi \rightarrow [\alpha]\phi$. ($\mathsf{Hist}_2$) is similar to the first of the Segerberg axioms (Harel 1984).

Axioms (NF_{Bel_i}) and (NL_{Bel_i}) can be put together into the single $(\neg[\alpha]\bot \wedge \neg Bel_i[\alpha]\bot) \rightarrow ([\alpha]Bel_i\phi \leftrightarrow Bel_i[\alpha]\phi)$. Equivalences of this kind have been called successor state axioms for belief in (Scherl & Levesque 1993).

(NF_{Choice_i}) and (NL_{Choice_i}) are their analogues for choice. Such axioms for choice have not been studied before.

(Inv_{Bel_i}) is a subjective version of a successor state axiom for belief and time. (Inv_{Choice_i}) is a similar axiom for choice and time. As far as we know they have not been studied before either.

From (NF_{Bel_i}) it follows that

$$Bel_i\Box\phi \wedge \neg Bel_i[\alpha]\bot \rightarrow [\alpha]Bel_i\Box\phi,$$

i.e. beliefs about invariants persist as long as there are no surprises.

From (Inv_{Bel_i}) it can be deduced in KD45 that

$$Bel_i\Box\phi \leftrightarrow Bel_i\Box Bel_i\phi$$

i.e. if i believes ϕ to be an invariant then he believes that he will always be aware of ϕ.

Moreover,

$$Bel_i\Box(Bel_i\phi \rightarrow \phi)$$
$$Bel_i\Diamond Bel_i\phi \rightarrow Bel_i\Diamond\phi$$
$$Choice_i\Diamond Bel_i\phi \rightarrow Choice_i\Diamond\phi$$

are valid.

The other way round, $Bel_i\Diamond\phi \rightarrow Bel_i\Diamond Bel_i\phi$ and $Choice_i\Diamond\phi \rightarrow Choice_i\ \Diamond Bel_i\phi$ should not hold. Here is an example illustrating that, inspired by Heisenberg's uncertainty principle. Let p mean that some electron is in a particular place. Suppose you believe that it will eventually be in that place: $Bel_i\Diamond p$. According to Heisenberg it is impossible to know that at the same point in time: $\Box\neg Bel_ip$. Now if we suppose that i is aware of that principle, we obtain $Bel_i\neg\Diamond Bel_ip$.

A similar argument can be made against $Choice_i\Diamond\phi \rightarrow Choice_i\Diamond Bel_i\phi$. This is opposed to Sadek and colleagues' approach (Sadek 1992; Bretier 1995; Louis 2003), where the principle $Choice_i\Diamond\phi \rightarrow Choice_i\Diamond Bel_i\phi$ is accepted.

We call ABC *logic* the logic thus axiomatized, and write $\vdash_{ABC}\phi$ if ϕ is a theorem of ABC.

THEOREM. $\models_{ABC}\phi$ iff $\vdash_{ABC}\phi$.

It is a routine task to check that all the axioms correspond to their semantic counterparts. It is routine, too, to check that all of our axioms are in the Sahlqvist class, for which a general completeness result exists (Sahlqvist 1975; Blackburn, de Rijke, & Venema 2001).

We conjecture that Marx's proof (1999) of decidability and EXPSPACE complexity of the problem of satisfiability in the product logic S5×K extends straightforwardly to ABC logic in the case of a single agent.[9]

In the rest of the paper, we apply ABC logic to investigate the notions of achievement goal, persistent goal, and intention.

Achievement goals

C&L view goals and intentions as particular future-oriented choices which take the form $Choice_i\Diamond\phi$.

If ϕ is already believed to be true then there is no point in maintaining the goal or the intention that ϕ. C&L therefore concentrate on goals which require some change in order to make them true. Basically such goals are of the form $Choice_i\Diamond\phi\wedge\neg\psi$, where ψ is a condition triggering the abandonment of the goal.

Which forms do ϕ and ψ take? First of all ϕ and ψ should be equivalent: when ϕ obtains then the goal can be abandoned, and whenever the goal is abandoned then ϕ holds.

[8] for example for $[\alpha]$:

from $\phi \leftrightarrow \psi$ infer $[\alpha]\phi \leftrightarrow [\alpha]\psi$ ($\mathsf{RE}_{[\alpha]}$)
$[\alpha](\phi \wedge \psi) \rightarrow [\alpha]\phi \wedge [\alpha]\psi$ ($\mathsf{M}_{[\alpha]}$)
$[\alpha]\phi \wedge [\alpha]\psi \rightarrow [\alpha](\phi \wedge \psi)$ ($\mathsf{C}_{[\alpha]}$)
$[\alpha]\top$ ($\mathsf{N}_{[\alpha]}$)

[9] We are indebted to Maarten Marx for pointing this out.

(This is at least expected by i.) Second, ψ should not be factual, but rather about i's mental state: else the agent has no means to decide when to abandon his goal. Hence achievement goals take the following form.

DEFINITION. Agent i has the achievement goal that ϕ if (1) in his preferred worlds ϕ is believed later and (2) i does not believe ϕ:

$$AGoal_i\phi \stackrel{\text{def}}{=} Choice_i\Diamond Bel_i\phi \wedge \neg Bel_i\phi \quad (\mathsf{Def}_{AGoal_i})$$

The only basic modal principle our definition of achievement goals validates is

$$\frac{\phi \leftrightarrow \psi}{AGoal_i\phi \leftrightarrow AGoal_i\psi}.$$

For the rest, just as in the C&L account none of the standard principles is valid.

The so-called *side effect problem* is to avoid to systematically adopt the consequences of our goals. Formally $AGoal_i\phi \wedge Bel_i(\phi \rightarrow \psi) \rightarrow AGoal_i\psi$ should not be valid. Just as for C&L, this formula is not valid in ABC logic. Even if we strengthen the condition $Bel_i(\phi \rightarrow \psi)$ in various ways, $AGoal_i\phi$ does not imply $AGoal_i\psi$. The reason is that the side effect might be believed, which makes that ψ cannot be an achievement goal. And just as C&L, if we add the condition $\neg Bel_i\psi$ then we validate

$$AGoal_i\phi \wedge Bel_i\Box(\phi \rightarrow \psi) \wedge \neg Bel_i\psi \rightarrow AGoal_i\psi.$$

(The proof makes use of the Axiom (Inv_{Bel_i}).) We also validate and the inference rule

$$\frac{\phi \rightarrow \psi}{AGoal_i\phi \wedge Bel_i \wedge \neg Bel_i\psi \rightarrow AGoal_i\psi}.$$

Finally, the valid equivalences

$$AGoal_i\phi \leftrightarrow Bel_iAGoal_i\phi$$

and

$$\neg AGoal_i\phi \leftrightarrow Bel_i\neg AGoal_i\phi$$

express that an agent is aware of his achievement goals. The equivalence

$$AGoal_i\phi \leftrightarrow AGoal_iBel_i\phi$$

is valid as well (while only the left-to-right direction is valid for C&L).

Comparison with C&L

C&L's original definition of achievement goals is

$$AGoal_i^{CL}\phi \stackrel{\text{def}}{=} Choice_i\Diamond\phi \wedge Bel_i\neg\phi.$$

THEOREM. $AGoal_i\phi \leftrightarrow AGoal_i^{CL}Bel_i\phi$.

This can be proved using introspection properties of belief.

C&L satisfy Axiom D: $\neg(AGoal_i\phi \wedge AGoal_i\neg\phi)$, while we do not.[10] Thus, while an agent's choices are consistent, his achievement goals are not necessarily so. This can be justified by the same temporal considerations that lead to rejection of axiom C: i might want ϕ to be true at some point in the future, and ϕ to be false at some other point in the future. But note that $AGoal_i\Box\phi \wedge AGoal_i\Box\neg\phi$ is unsatisfiable due to the confluence of time.

In their definition, C&L stipulate that i should believe ϕ is false. We have preferred the weaker $\neg Bel_i\phi$ because it is more natural: in general goals are abandoned only when they are believed to be true, and therefore absence of belief is sufficient to maintain the goal (but see our Byzantine example below for a counterexample).

C&L only require $Choice_i\Diamond\phi$. We have seen in the previous section that $Choice_i\Diamond Bel_i\phi \rightarrow Choice_i\Diamond\phi$ is a theorem. We have also said there that the other sense of the implication should not hold. So let us consider a situation where $Choice_i\Diamond\phi \wedge \neg Choice_i\Diamond Bel_i\phi$ holds. The following example seems to motivate the need for achievement goals in C&L's sense.

Let r mean that a message of i has been received by j, and let i believe initially that j has not received the message yet. Suppose we are in a Byzantine-generals-style scenario where i is not guaranteed that his message will eventually be received by j, and where i believes that in any case he will never know whether j received the message or not. (In the original scenario it is just possible for i that he will never know.) Hence we have $Bel_i\neg r \wedge Choice_i\Diamond r \wedge Bel_i\Box\neg BelIf_ir$. From the latter it follows that $\neg Choice_i\Diamond Bel_ir$. In summary, we have $Bel_i\neg r \wedge AGoal_i^{CL}r \wedge \neg AGoal_ir$.

Now in such a context it seems reasonable that i acts by nevertheless posting the message. C&L can account for this case by stating $AGoal_i^{CL}r$. What would be i's achievement goal in our account? We argue that in the example i has the achievement goal that $\neg Bel_i\neg r$: such an achievement goal can first motivate i to post the message, and then trigger abandonment (say after the time period i esteems necessary for the message travelling under favorable conditions). Note that $AGoal_i\neg Bel_i\neg r$ is consistent with the scenario description.

Consider another example where there is only one action of toggling a switch, and suppose that in the initial world $w_0 \models \neg Bel_iLight \wedge \neg Bel_i\neg Light$, i.e. i ignores whether the light is on or off: for i there is at least one possible world where $Light$ holds, and there is at least one possible world where $\neg Light$ holds. As toggling is the only available action we have $w_0 \models Bel_i\Box(\neg Bel_iLight \wedge \neg Bel_i\neg Light)$, i.e. i believes he will always ignore whether the light is on or off. According to C&L agent i can nevertheless have the

[10] As C&L's admit, this is 'for the wrong reasons': their stronger definition of achievement goals is responsible for $AGoal_i\phi \rightarrow Bel_i\neg\phi$, which warrants axiom D for $AGoal_i$. Note that they do not validate the stronger but equally intuitive principle $\frac{\neg(\phi\wedge\psi)}{\neg(AGoal_i\phi\wedge AGoal_i\psi)}$. Apparently this has not been noted in the literature.

achievement goal $AGoal_i^{CL} Light$ in w_0, while he cannot have such a goal with our definition. Thus i is aware that he will never be able to abandon his goal that $Light$ in the expected way, viz. by coming to believe that $Light$.

Persistent goals

C&L have defined persistent goals to be achievement goals that are kept until they are achieved, or are abandoned for some other reasons. We can show that persistence can be *deduced* from our no forgetting principle for choice as long as the event is not unwanted:

THEOREM. $\models_{ABC} \ (AGoal_i\phi \wedge \neg Choice_i[\alpha]\bot) \ \rightarrow [\alpha](AGoal_i\phi \vee Bel_i\phi)$

PROOF. We prove $\neg Bel_i\phi \wedge Choice_i\Diamond Bel_i\phi \ \rightarrow Choice_i[\alpha]\bot \vee [\alpha]Choice_i\Diamond Bel_i\phi$. This can be deduced from (NL$_{Choice_i}$), (Hist2), (Inc$_{Choice_i}$) as follows.

First, axiom (Hist2) tells us that

$$\Diamond Bel_i\phi \rightarrow (Bel_i\phi \vee [\alpha]\Diamond Bel_i\phi)$$

for any action α. Therefore

$$Choice_i\Diamond Bel_i\phi \rightarrow Choice_i(Bel_i\phi \vee [\alpha]\Diamond Bel_i\phi).$$

As by (5$_{Bel_i}$) and (Inc$_{Choice_i}$) we have

$$\neg Bel_i\phi \rightarrow Choice_i\neg Bel_i\phi,$$

the left hand side implies

$$Choice_i[\alpha]\Diamond Bel_i\phi.$$

From that we get with (NL$_{Choice_i}$) that

$$Choice_i[\alpha]\bot \vee [\alpha]Choice_i\Diamond Bel_i\phi.$$

■

We inherit the properties of achievement goals concerning logical principles, the side effect problem, and persistence.

Comparison with C&L

C&L's original definition is that a persistent goal that ϕ is an achievement goal that ϕ that can only be abandoned if

1. ϕ is achieved, or
2. the agent learns that ϕ can never be achieved, or
3. for some other reason.

This leads to their principle

$PGoal_i\phi \ \rightarrow \ [\alpha](PGoal_i\phi \vee Bel_i\phi \vee Bel_i\Box\neg\phi \vee \psi)$,

where ψ is an unspecified condition accounting for case (3). Our theorem makes (3) more precise by identifying it with the occurrence of an unwanted event, which is the only case when achievement goals have to be revised.[11] Indeed, the theorem tells us that C&L's case (2) is excluded when $\neg Choice_i[\alpha]\bot$ holds: in this case we are guaranteed that i will not learn through α that ϕ will be false henceforth. Given our hypothesis that events are uninformative, this is as it should be.

[11] In the case where i is the agent of α (noted $i:\alpha$) one might reasonably suppose that $Choice_i[i{:}\alpha]\bot \ \rightarrow \ [i{:}\alpha]\bot$, i.e. there are no such unwanted action occurrences. We then get unconditioned persistence of achievement goals: $AGoal_i\phi \rightarrow [i{:}\alpha](AGoal_i\phi \vee Bel_i\phi)$. This is related to intentional actions as discussed in C&L's (1990a, section 4.2.1), where moreover $Bel_i[i{:}\alpha]\bot \vee Bel_i\neg[i{:}\alpha]\bot$ is assumed. We just note that such principles are of the Sahlqvist type, and can be added to ABC logic without harm.

Intentions

C&L have distinguished intentions-to-do and intentions-to-be. We here only consider the latter, which, following Bratman, C&L have defined as particular persistent goals: the agent must be committed to achieve the goal, in the sense that he must believe that he will perform an action which will lead to the goal.

DEFINITION. Agent i has the intention that ϕ if (1) i has the achievement goal that ϕ, and (2) i does not believe $Bel_i\phi$ will obtain anyway:

$$Int_i\phi \stackrel{\text{def}}{=} AGoal_i\phi \wedge \neg Bel_i\Diamond Bel_i\phi \qquad (\text{Def}_{Int_i})$$

Hence intentions are achievement goals which do not automatically obtain in the future. As $\neg Bel_i\Diamond Bel_i\phi$ implies $\neg Bel_i\phi$, it follows that $Int_i\phi \ \leftrightarrow \ Choice_i\Diamond Bel_i\phi \wedge \neg Bel_i\Diamond Bel_i\phi$. If not explicitly, this implicitly links i's intending that ϕ to i's choosing actions that get him closer to ϕ: $Int_i\phi$ triggers i's planning for ϕ. Therefore it seems justified to say that our definition captures the spirit of Bratman's intentions.

What is the status of achievement goals when $Bel_i\Diamond Bel_i\phi$ holds? In this case, $AGoal_i\phi \wedge Bel_i\Diamond Bel_i\phi$ is equivalent to $Bel_i\Diamond Bel_i\phi \wedge \neg Bel_i\phi$: i believes ϕ will be achieved in the future, no matter what continuation of his possible histories occurs. Then according to our definition i has to abandon $Int_i\phi$ at w_1. This is reminiscent of McDermott's Little Nell example: suppose that i intends that ϕ at w_0, and that i successfully plans and acts in a way such that later on at w_1 he is sure ϕ will be achieved in the future, i.e. $Bel_i\Diamond Bel_i\phi$ holds at w_1. According to McDermott i then abandons his intention that ϕ too early, and will never achieve ϕ. We believe the problem can be solved by separating planning-oriented (future-oriented) intention from intention-in-action: at w_1 agent i switches from the planning-oriented intention $Int_i\phi$ to the intention-in-action to execute the plan (alias complex action) which he believes ensures that ϕ will obtain. i will stick to this plan from w_1 on and as long as no unforeseen events occur.[12]

Again, we inherit the properties of achievement goals concerning logical principles, the side effect problem, and in particular persistence:

THEOREM. $\models_{ABC} \ (Int_i\phi \wedge \neg Choice_i[\alpha]\bot) \ \rightarrow [\alpha](Int_i\phi \vee Bel_i\Diamond Bel_i\phi)$

PROOF. The theorem of the previous section establishing that achievement goals are also persistence goals, a look at the proof tells us that

$$(AGoal_i\phi \wedge \neg Choice_i[\alpha]\bot) \rightarrow [\alpha]Choice_i\Diamond Bel_i\phi$$

Therefore by classical principles

$$(AGoal_i\phi \wedge \neg Choice_i[\alpha]\bot) \rightarrow [\alpha]((Choice_i\Diamond Bel_i\phi \wedge \neg Bel_i\Diamond Bel_i\phi) \vee Bel_i\Diamond Bel_i\phi)$$

[12] We could pursue this and define future-directed intention-to-do α as $Choice_i\Diamond\langle i{:}\alpha\rangle\top$.

from which the present theorem follows by the definition of intention. ■

Hence intentions persist as long as there are no unwanted action occurrences.

Comparison with C&L

Our definition of $Int_i\phi$ differs from C&L's in a fundamental way because it does not mention actions: C&L basically stipulate that in every preferred history there must be some action α whose author is i and which brings about ϕ.

Using quantification over actions this could be approximated by:

$$Int_i^{CL}\phi \stackrel{\text{def}}{=} \neg Bel_i\phi \wedge Choice_i\Diamond\exists i{:}\alpha\langle i{:}\alpha\rangle Bel_i\phi.$$

But as pointed out by Sadek (2000) and Bretier (1995), such a definition is too strong in particular in cooperative contexts, where it often suffices for i to *trigger* actions of some other agent j which will achieve the goal. They have advocated a correction, which we roughly approximate here by:

$$Int_i^{S}\phi \stackrel{\text{def}}{=} \neg Bel_i\phi \wedge Choice_i\Diamond Bel_i\phi \wedge Choice_i\forall i{:}\alpha(Bel_i\langle i{:}\alpha\rangle\Diamond Bel_i\phi \rightarrow Choice_i\Diamond\langle i{:}\alpha\rangle\top).$$

Again, this is too strong: my intention to go to Vancouver in june here would force me to choose the action of hiring an aircraft. In another sense, both C&L's and Sadek's definitions are too weak because they lack a causal connection between the action and the goal: basically they entitle me to entertain the intention that it be sunny in Vancouver in june if each of my preferred histories has some action of mine leading to a state where this holds.

As our definition of intention does not mention events at all, this example also illustrates that our definition is also too weak in this respect.

Conclusion

We have integrated action, time, belief, and choice in a simple propositional modal logic that is sound, complete and decidable, and which we think provides the basic framework for the logical analysis of interaction. We have shown how different notions of goal and intention can be expressed in it, and have identified the conditions under which such motivational attitudes persist.

Although Cohen and Levesque's papers are standard references, to the best of our knowledge such a simplification has never been undertaken. Our completeness, decidability and complexity results pave the way for methods of mechanical deduction.

In ABC logic we have also in part solved the frame problem for belief and intention. While the frame problem for belief has been investigated extensively in the literature, there is not too much work in the literature on the frame problem for intentions, and the only references we are aware of are (Shapiro & Lespérance 2000; Shapiro, Lespérance, & Levesque 1997; 1998). These accounts are preliminary, in particular they lead to fanatic agents.

What is lacking for a comprehensive solution to the frame problem for intention is the integration of belief and choice revision (sometimes called intention reconsideration in agent theories (Thomason 2000; Schut & Wooldridge 2001)). We leave this important issue to future work.

What remains also to be addressed is the question of how intentions lead to actions. This is is the topic of plan generation, which still has to be integrated in our logic.

References

Appelt, D., and Konolige, K. 1989. A nonmonotonic logic for reasoning about speech acts and belief revision. In Reinfrank, M.; de Kleer, J.; Ginsberg, M.; and Sandewall, E., eds., *Proc. 2nd Int. Workshop on Non-monotonic Reasoning*, number 346 in LNAI, 164–175. Springer Verlag.

Bacchus, F.; Halpern, J.; and Levesque, H. 1995. Reasoning about noisy sensors in the situation calculus. In *Proc. 14th Int. Joint Conf. on Artificial Intelligence (IJCAI'95)*, 1933–1940.

Bacchus, F.; Halpern, J.; and Levesque, H. 1999. Reasoning about noisy sensors in the situation calculus. *Artificial Intelligence* 111:131–169.

Baltag, A.; Moss, L. S.; and Solecki, S. 1998. The logic of public announcements, common knowledge, and private suspicions. In *Proc. TARK'98*, 43–56. Morgan Kaufmann.

Baltag, A. 2000. A logic of epistemic actions. Technical report, CWI. `http://www.cwi.nl/~abaltag/papers.html`.

Blackburn, P.; de Rijke, M.; and Venema, Y. 2001. *Modal Logic*. Cambridge Tracts in Theoretical Computer Science. Cambridge University Press.

Bratman, M. E. 1987. *Intentions, plans, and practical reason*. Harvard University Press, MA.

Bratman, M. E. 1990. What is intention? In Cohen, P. R.; Morgan, J.; and Pollack, M. E., eds., *Intentions in Communication*. Cambridge, MA: MIT Press. chapter 2, 15–31.

Bretier, P. 1995. *La communication orale coopérative : contribution à la modélisation logique et à la mise en œuvre d'un agent rationnel dialoguant*. Ph.D. Dissertation, Université Paris Nord, Paris, France.

Cohen, P. R., and Levesque, H. J. 1990a. Intention is choice with commitment. *Artificial Intelligence J.* 42(2–3):213–261.

Cohen, P. R., and Levesque, H. J. 1990b. Persistence, intentions, and commitment. In Cohen, P. R.; Morgan, J.; and Pollack, M. E., eds., *Intentions in Communication*. Cambridge, MA: MIT Press. chapter 3, 33–69.

Harel, D. 1984. Dynamic logic. In Gabbay, D. M., and Günthner, F., eds., *Handbook of Philosophical Logic*, volume II. D. Reidel, Dordrecht. 497–604.

Herzig, A., and Longin, D. 2000. Belief dynamics in cooperative dialogues. *J. of Semantics* 17(2). vol. published in 2001.

Herzig, A., and Longin, D. 2002. Sensing and revision in a modal logic of belief and action. In van Harmelen, F., ed., *Proc. ECAI2002*, 307–311. IOS Press.

Konolige, K., and Pollack, M. E. 1993. A representationalist theory of intention. In *Proc. 13th Int. Joint Conf. on Artificial Intelligence (IJCAI'93)*, 390–395. Chambery, France: Morgan Kaufmann.

Louis, V. 2003. *Conception et mise en œuvre de modles formels du calcul de plans d'action complexes par un agent rationnel dialoguant*. Ph.D. Dissertation, Université de Caen, France.

Marx, M. 1999. Complexity of products of modal logics. *J. of Logic and Computation* 9(2):221–238.

Moore, R. C. 1985. A formal theory of knowledge and action. In Hobbs, J., and Moore, R., eds., *Formal Theories of the Commonsense World*. Norwood, NJ: Ablex. 319–358.

Perrault, C. R. 1990. An application of default logic to speech act theory. In Cohen, P. R.; Morgan, J.; and Pollack, M. E., eds., *Intentions in Communication*. Cambridge, MA: MIT Press. chapter 9, 161–185.

Rao, A. S., and Georgeff, M. P. 1991. Modeling rational agents within a BDI-architecture. In Allen, J. A.; Fikes, R.; and Sandewall, E., eds., *Proc. 3rd Int. Conf. on Knowledge Representation and Reasoning (KR'91)*, 473–484. San Mateo, CA: Morgan Kaufmann.

Rao, A. S., and Georgeff, M. P. 1992. An abstract architecture for rational agents. In Nebel, B.; Rich, C.; and Swartout, W., eds., *Proc. 4th Int. Conf. on Knowledge Representation and Reasoning (KR'92)*, 439–449. Cambridge, Massachusetts: Morgan Kaufmann.

Sadek, M. D. 1992. A study in the logic of intention. In Nebel, B.; Rich, C.; and Swartout, W., eds., *Proc. 4th Int. Conf. on Knowledge Representation and Reasoning (KR'92)*, 462–473. Cambridge, Massachusetts: Morgan Kaufmann.

Sadek, M. D. 2000. Dialogue acts are rational plans. In Taylor, M.; Nel, F.; and Bouwhuis, D., eds., *The structure of mutimodal dialogue*, 167–188. Philadelphia/Amsterdam: John Benjamins publishing company. From ESCA/ETRW, Workshop on The Structure of Multimodal Dialogue (Venaco II), 1991.

Sahlqvist, H. 1975. Completeness and correspondence in the first and second order semantics for modal logics. In Kanger, S., ed., *Proc. 3rd Scandinavian Logic Symposium 1973*, number 82 in Studies in Logic. North Holland.

Scherl, R., and Levesque, H. J. 1993. The frame problem and knowledge producing actions. In *Proc. Nat. Conf. on AI (AAAI'93)*, 689–695. AAAI Press.

Scherl, R., and Levesque, H. J. 2003. The frame problem and knowledge producing actions. *Artificial Intelligence* 144(1-2).

Schut, M., and Wooldridge, M. 2001. Principles of intention reconsideration. In *Proc. AGENTS'01*. ACM Press.

Shapiro, S., and Lespérance, Y. 2000. Modeling multiagent systems with the cognitive agents specification language - a feature interaction resolution application. In Castelfranchi, C., and Lespérance, Y., eds., *Intelligent Agents Vol. VII - Proc. 2000 Workshop on Agent Theories, Architectures, and Languages (ATAL-2000)*. Springer-Verlag.

Shapiro, S.; Pagnucco, M.; Lespérance, Y.; and Levesque, H. J. 2000. Iterated belief change in the situation calculus. In *Proc. KR2000*, 527–538.

Shapiro, S.; Lespérance, Y.; and Levesque, H. J. 1997. Specifying communicative multi-agent systems with ConGolog. In *Working notes of the AAAI fall symposium on Communicative Action in Humans and Machines*, 75–82. AAAI Press.

Shapiro, S.; Lespérance, Y.; and Levesque, H. J. 1998. Specifying communicative multi-agent systems. In Wobcke, W.; Pagnucco, M.; and Zhang, C., eds., *Agents and Multi-Agent Systems - Formalisms, Methodologies, and Applications*, 1–14. Springer-Verlag, LNAI 1441.

Thomason, R. H. 2000. Desires and defaults: A framework for planning with inferred goals. In *Proc. of the Seventh International Conference Knowledge Representation and Reasoning (KR'2000)*, 702–713. Morgan Kaufmann Publishers.

Wooldridge, M. 2000. *Reasonning about Rational Agent*. Cambridge, Massachusetts: MIT Press.

Sensing and revision in a modal logic of belief and action

Andreas Herzig[1] and **Dominique Longin**[2]

Abstract. We propose a modal logic of belief and actions, where action might be nondeterministic, and there might be misperception.

The agent must be able to revise his beliefs, because (contrarily to knowledge) observations might be inconsistent with his beliefs. We propose a new solution in terms of successor state axioms, which does not resort to orderings of plausibility. Our solution allows for regression in the case of deterministic actions.

1 Introduction

Since the beginning of the 90ies, solutions to the frame problem have been extended to cover perception [17, 13, 20, 23]. In these approaches perception has been analyzed in terms of actions. To such perception actions one can opposes *uninformative actions*, which are actions whose outcome is not perceived by the agent. (When the agent learns that such an action has occurred, he is nevertheless able to predict its results according to the action laws.) It is noted in several places (e.g. [17],[21, footnote 10],[10]) that actions can be analysed as a sequence of uninformative actions and perception actions. For example, the action of tossing a coin can be decomposed into the uninformative action of tossing without observing the result – eyes shut –, followed by the perception action of checking the result. The most important class of uninformative actions are ontic actions (physical actions), which are actions that can be described without referring to belief.

Perception actions are reduced to actions of observing that some proposition is true: that the light in some room is on, that tossing a coin resulted in heads, etc. We call such actions *observation actions*. We suppose that they do not change the environment, but only the agent's mental state. (For the sake of simplicity we suppose that there is only one agent.)

When reasoning about observations one has to distinguish what is true from what is believed by an agent: it might be the case that some proposition A is true, but the agent is not aware of it. Therefore we suppose that in every situation (alias possible world) w the agent entertains a set of beliefs $B(w)$.

Now suppose some action a occurs, resulting in a new situation w'. What is $B(w')$ like? If a is uninformative, then $B(w')$ should only depend on $B(w)$, a and the action laws: the agent predicts the result of a using the action laws for a. Indeed, apart from the mere execution of a the agent should learn nothing about a's particular effects that hold in w'.

According to this account, observation actions are uninformative: to learn that the observation of A has occurred means to learn that A. All the relevant information is thus encoded in the notification of the action occurrence. Then to take into account a amounts to incorporate A into $B(w)$.[3]

Take the action of tossing a coin. When the agent is notified the occurrence of the tossing action, as he cannot observe the effects of toss, he predicts them in an a priori way, according to his mental state and the action laws. The agent can thus be said to "mentally execute" toss. Hence afterwards he believes that Heads$\vee$Tails holds, but neither believes Heads nor Tails. When the agent subsequently learns that the coin fell heads (being notified that Heads has been observed) then he moves to believing that Heads.

In consequence we can restrict our attention to uninformative actions. We focus on the following type of scenarios:

- in a given situation w the agent entertains a set of beliefs $B(w)$;
- some action a occurs, resulting in a new situation w';
- the agent is notified that some a' has occurred (where $a' \neq a$ if there is misperception);
- the agent does not learn which of the effects of a hold in w;
- the agent takes into account the occurrence of a' by appropriately changing $B(w)$, and forms the new set $B(w')$ that he holds in w'.

We have thus generalized our account to allow for misperception.

Most of the approaches in the reasoning about actions domain are formulated in terms of a modal operator of knowledge [14, 17, 20, 23, 21, 10]. Knowledge being viewed as true belief, in such approaches *surprises* are impossible: if an agent knows that A then A must be true; as observations don't change the environment, A still holds after any observation; hence $\neg A$ can only be observed if there is misperception, but in this case the agent realizes that, and immediately rejects the input.

It follows that two operations are enough to implement knowledge change: updates *à la* Katsuno-Mendelzon (KM-updates) [11] to take into account uninformative actions, and expansions *à la* Alchourrón-Gärdenfors-Makinson (AGM-expansions) [1] to take into account perception actions.

The picture is different in the case of belief change, because beliefs can be contradicted by observations: I believe that I have a coin in my pocket, but on checking I find out I don't; I believe my watch is waterproof but when trying it out it isn't, etc. It is non-trivial to extend the above solutions to handle such examples. Expansion operations do not suffice: we need belief *revision* operations *à la* AGM.

More generally, the problem arises as soon as the agent believes some action is inexecutable and nevertheless learns that it has occurred. In this case the agent must first revise his current beliefs by the preconditions of the action, and then apply the action laws associated to the action.

Many authors raise the issue and are aware of the difficulties (e.g.

[1] Institut de Recherche en Informatique de Toulouse, France, email: Andreas.herzig@irit.fr

[2] Laboratoire Travail et Cognition, Université Toulouse-le-Mirail, France, email: longin@univ-tlse2.fr

[3] This does not hold for all perception actions, such as testing if a proposition is true. Such tests can nevertheless be reduced to uninformative actions, see Sect. 2.5.

[17, 6]), but the only proposal up to now is that of Shapiro *et al*. [21], which is based on orderings of plausibility.

In the sequel we shall do without such a device. Our approach is characterized by the following hypotheses:

(H1) All atomic actions are either ontic or observation actions.
(H2) Actions might be non-deterministic.
(H3) There might be misperception and non-perception of action occurrences.
(H4) Uninformative actions do not affect the agent's cognition. Hence we exclude actions such as modifying the agent's memory.
(H5) The action laws are known by the agent.

In Sect. 2 we introduce a logic for belief and action, which is similar to Segerberg's DDL [19, 18]. Then we focus on uninformative and observation actions, and show how updating and revision can be done (Sect. 3, Sect. 4, Sect. 5). Finally we discuss related work (Sect. 6).

2 Dynamic Doxastic Logic

2.1 Belief

We suppose that our language contains a modal operator of belief Bel. The formula $BelA$ is read "the agent believes that A". $BelIf\,A$ is defined as $BelIf\,A \stackrel{\text{def.}}{=} BelA \vee Bel\neg A$, and can be read "the agent believes A or believes $\neg A$", or more shortly "the agent knows[4] whether A is true or not".

We adopt the modal logic KD45 as the logic of belief, i.e. we suppose agents do not entertain inconsistent beliefs, and are aware of their beliefs and disbeliefs.

2.2 Actions

We use a simple version of PDL [8] to speak about actions. Actions are noted a, a', b, ... The empty action is noted λ. To each action a there is associated a modal operator $After_a$. The formula $After_a A$ reads "A is true after a". $After_a \bot$ expresses that a is inexecutable. An example of a formula involving belief and action is $Bel\neg After_a \bot \wedge After_a \bot$, expressing that the agent believes that a can be executed, while this is not the case. The operator $Feasible_a$ is introduced as an abbreviation: $Feasible_a A \stackrel{\text{def.}}{=} \neg After_a \neg A$. $Feasible_a \top$ expresses that a is executable.

We adopt the standard axiomatics of PDL, which for our fragment is nothing but the multimodal logic K. ($After_a$ corresponds to the Dynamic Logic operator $[a]$, and $Feasible_a$ to $\langle a \rangle$.)

2.3 Possible worlds semantics

We adopt the standard possible worlds semantics, with models having a set of situations W, and accessibility relations R_{Bel} and R_a respectively associated to the modal operators Bel and $After_a$. We view the belief state of an agent in a given situation w as a set of possible worlds $R_{Bel}(w) = \{v : wR_{Bel}v\}$, and $v \in R_{Bel}(w)$ means that the situation v is compatible with the agent's beliefs. $R_a(w) = \{w' : wR_a w'\}$ is the set of possible results w' of action a when applied in w.

R_{Bel} is reflexive, transitive and euclidean. In a situation $w \in W$, the set of situations $R_{Bel}(w) = \{u \in W : wR_{Bel}u\}$ is called the *belief state* of the agent in w.

[4] We use the term "knows" here because "the agent believes whether A" sounds odd.

Actions might be nondeterministic (because the R_a are not necessarily functions), and they might be inexecutable (when there is no w' such that $wR_a w'$).

2.4 Misperception

In most of the related approaches [17, 21, 23] it is supposed that actions are public: when a occurs then its occurrence is correctly notified to the agent. This means that (1) if an agent believes that some action a occurred then a indeed occurred (correctness), and (2) if a occurred then the agent believes a occurred (completeness).

We suppose here that the agent might instead perceive some other action b. The atomic proposition $perc(a, b)$ expresses that the occurrence of a has been notified as the occurrence of b by the agent. Note that setting b to the empty action allows to simulate the case where the agent is unaware that an action has occurred. The other way round, setting a to the empty action allows to simulate illusions.

Note that in terms of Sandewall's systematic approach [16], most of the approaches in reasoning about actions suppose that knowledge is explicit, accurate and correct (Sandewall's class $\mathcal{K}$). Hence there is no misperception or illusion.

Several approaches to misperception exist in the literature, e.g. [2, 4]. In [4] a classification is given. Among the three cases there, we can account here for the case where observations do not agree with the effects that actions are supposed to have, and the case where new observations indicate unpredicted change (conflicting with the principle of inertia). Among the different revision strategies that are discussed, the one we adopt here is that of preferring the last observation when constructing the new belief state.

2.5 Observation actions

We note $observe(A)$ the action of observing that A. Observation actions can be characterised by the following *logical axioms*[5] (see [9, 10]).

$$A \rightarrow Feasible_{observe(A)} \top \qquad \text{(TestAct}_1\text{)}$$
$$\neg A \rightarrow After_{observe(A)} \bot \qquad \text{(TestAct}_2\text{)}$$
$$After_{observe(A)} A \qquad \text{(TestAct}_3\text{)}$$
$$C \rightarrow After_{observe(A)} C \quad \text{if} \quad C \text{ is objective} \qquad \text{(TestAct}_4\text{)}$$
$$Feasible_{observe(A)} C \rightarrow After_{observe(A)} C \qquad \text{(TestAct}_5\text{)}$$

An *objective formula* is a formula without occurrences of the doxastic modal operator Bel. The first two axioms together say that $observe(A)$ is executable iff A is true. Therefore learning that $observe(A)$ has been executed amounts to learning that A. (TestAct$_3$) says that A holds after observing that A. Together with the more general principles of sections Sect. 3 and Sect. 4 it will guarantee that observing that A leads to believing that A. (TestAct$_4$) expresses that observation actions are perception actions, and the last says they are deterministic.[6]

Observation actions behave as expansions in the AGM-theory: $observe(A)$ makes shrink the belief state by 'throwing out' those possible situations where A is false.

[5] Note that logical axioms are known by the agent. (This is obtained by the necessitation rule of KD45.)

[6] $observe(A)$ is similar to the PDL test "$A?$". The difference is that for the latter $After_{A?} C$ is defined as $A \rightarrow C$. Hence such tests validate $B \rightarrow After_{A?} B$ for *every* formula B. However, consider $B = \neg BelIf\,A$: intuitively, the formula $\neg BelIf\,A \rightarrow After_{A?} \neg BelIf\,A$ should not be valid. Therefore such a principle must be restricted to objective B's, which is what we did here.

Nondeterministic composition of $observe(A)$ and $observe(\neg A)$ can 'simulate' the action $testIf(A)$ of testing-if A: testing if the coin is heads amounts to nondeterministically choose between $observe(A)$ and $observe(\neg A)$ and execute the chosen action. Therefore testing-if can be viewed as an abbreviation of testing-that:

$$After_{testIf(A)}B \stackrel{\text{def.}}{=} After_{observe(A)}B \wedge After_{observe(\neg A)}B$$

It can then be proved that $Feasible_{testIf(A)}\top$ holds, as well as $After_{testIf(A)}BelIf A$ and $A \rightarrow After_{testIf(A)}BelA$. Note that while $observe(A)$ is an uninformative action, $testIf(A)$ is not.

2.6 Ontic actions

We suppose that to each ontic action a there is associated a set of *effect laws* and a set of *executability laws*. The former are of the form $A \rightarrow After_a C$ and the latter are of the form $A \rightarrow Feasible_\alpha \top$ where A and C must be *factual*, i.e. without any modal operator.

For example, the (ontic) toss action is executable if one has a coin: $HasCoin \rightarrow Feasible_{\mathsf{toss}}\top$, and has the effects $\mathsf{Heads} \vee \mathsf{Tails}$ and $\neg HasCoin$: $After_a((\mathsf{Heads} \vee \mathsf{Tails}) \wedge \neg HasCoin)$.

3 Updating

Semantically a (non-deterministic) action is a relation R_a between possible situations – alias possible worlds –, where $w' \in R_a(w)$ means that w' is a possible result of a when applied in w. We view the belief state of an agent in a given situation w as a set of possible worlds $R_{Bel}(w)$, and $v \in R_{Bel}(w)$ means that the situation v is compatible with the agent's beliefs. The occurrence of an action makes the current situation w evolve to a new situation $w' \in R_a(w)$. What can we say about $R_{Bel}(w')$, i.e. the agent's belief state at w'?

First of all, it might be the case that the agent is not notified of a, but some other action b, as expressed by the atomic proposition $perc(a, b)$. How should the agent take into account such an action occurrence b? Following Moore [14] and Scherl and Levesque [17], the agent's belief state $R_{Bel}(w')$ in w' results from applying action b to all possible worlds in $R_{Bel}(w)$ ("mentally executing b"), and collecting the resulting situations:

$$R_{Bel}(w') = \bigcup_{v \in R_{Bel}(w)} R_b(v)$$

This looks fine, but there is a problematic case here when $R_b(v) = \emptyset$ for every $v \in R_{Bel}(w)$: then $R_{Bel}(w') = \emptyset$, which would mean that the agent ends up with an inconsistent set of beliefs. This contradicts our hypothesis that beliefs are consistent (axiom D). Under such a proviso our axiom for updates is the generalization of the successor state axiom for knowledge of [17] to belief and non-deterministic actions:

$$(perc(a, b) \wedge \neg After_a\bot \wedge \neg BelAfter_b\bot) \rightarrow (Feasible_a BelA \leftrightarrow BelAfter_b A) \qquad (\text{SSA}_\text{I})$$

where a is an uninformative action. It says that the agent cannot observe anything after a is performed: indeed, for any formula A, if he cannot predict before a is performed that A will hold after a is performed, then he will not know A after a is performed.

Consider e.g. $a = \mathsf{toss}$, and suppose $perc(\mathsf{toss}, \mathsf{toss})$ and $HasCoin \wedge BelHasCoin$ hold. It follows from the executability laws for toss that $\neg After_{\mathsf{toss}}\bot$ and $\neg BelAfter_{\mathsf{toss}}\bot$ (the latter by necessitation and axiom D). We therefore have for $A = \mathsf{Heads}$: $Feasible_{\mathsf{toss}} Bel\mathsf{Heads} \leftrightarrow BelAfter_{\mathsf{toss}}\mathsf{Heads}$. From the left to the right, (SSA_I) expresses that for uninformative actions there are no a posteriori beliefs the agent didn't already hold a priori: if after some execution of toss the agent believes that the coin is heads (i.e. $Feasible_{\mathsf{toss}} Bel\mathsf{Heads}$) then – as he had no means to check whether Heads is true – he must have believed before tossing that the coin is biased, i.e. $BelAfter_{\mathsf{toss}}\mathsf{Heads}$. Reading the equivalence from the right to the left, consider $A = \neg BelIf\,\mathsf{Heads}$. ($\text{SSA}_\text{I}$) expresses that the agent's uncertainty about the nondeterministic result of toss is preserved through its execution: before executing toss the agent ignores whether heads or tails will result: $BelAfter_{\mathsf{toss}}\neg BelIf\,\mathsf{Heads}$ and this disbelief $\neg BelIf\,\mathsf{Heads}$ is preserved through the execution of toss: $Feasible_{\mathsf{toss}} Bel\neg BelIf\,\mathsf{Heads}$, which is equivalent to $Feasible_{\mathsf{toss}}\neg BelIf\,\mathsf{Heads}$.

To take another example, suppose $Bel(\neg\mathsf{Heads} \wedge \neg\mathsf{Tails})$ holds, and suppose that a toss-action takes place but the agent isn't notified. Hence $perc(a, \lambda)$ holds, and $After_{\mathsf{toss}} Bel(\neg\mathsf{Heads} \wedge \neg\mathsf{Tails})$.

As said in section 2, observations are uninformative actions, and (SSA_I) applies to observation actions, too: if after observing that Heads I believe that $\mathsf{Heads} \wedge A'$ for some A', and $observe(\mathsf{Heads})$ is executable, then I believe before $observe(\mathsf{Heads})$ that $\mathsf{Heads} \wedge A'$ will be true afterwards.

Remark 1 *The only case where the $\rightarrow$ direction of the equivalence in (SSA_I) cannot be accepted is when a erases all or part of the memory of the agent (e.g. taking off the batteries of a robot). The $\leftarrow$ direction is counter-intuitive only if the agent knows that a adds unjustified information to his memory. This is the case e.g. when he is hypnotized or takes drugs. We have excluded such extreme cases by hypothesis (H4).*

Finally, as announced in the end of section 2, from (SSA_I), (TestAct_3) and standard modal principles we can prove the following theorem:

Theorem 1 $After_{observe(A)} BelA$

4 Revising

4.1 Enabling actions

Suppose the agent believes a is inexecutable, and learns that a has nevertheless occurred. Axiom (SSA_I) says nothing about that case. Such surprising occurrences of actions are indeed problematic, because the agent is unable to just mentally execute a, and must first change his beliefs about a's preconditions.

We propose to formalize the operation of changing beliefs about preconditions by the mental execution of a particular ontic action whose effect is to make the executability preconditions of a true. This makes sense when applied to possible worlds: if an agent believes a to be inexecutable but nevertheless learns that a has happened, he adjusts each of his possible worlds w so as to enable executability of a in w.

Formally, we associate to every atomic action a an action $enable_a$, and we say that $enable_a$ makes a executable.

We postulate that $enable_a$ can occur in every situation. Hence we have the executability law

$$Feasible_{enable_a}\top \qquad (\text{Exec}_{enable_a})$$

This means that for every action there is at least one situation where it is executable. Note that this excludes from our actions the action $observe(\bot)$ which is never executable.

Let a be any ontic action. According to our definition, the set of executability laws for a has the form

$$\{A_1 \rightarrow \mathit{Feasible}_a\top, \ldots, A_n \rightarrow \mathit{Feasible}_a\top\}$$

As enable_a makes a executable, the set of executability laws for a determines the following effect law for enable_a:

$$\mathit{After}_{\mathit{enable}_a}(A_1 \vee \ldots \vee A_n) \qquad (\text{Eff}_{\mathit{enable}_a})$$

For example, for the toss action, if the agent believes there is no coin, and nevertheless learns that toss has been executed, then he enables toss in his possible worlds: $\mathit{After}_{\mathit{enable}_{\mathsf{toss}}}\mathit{HasCoin}$.

What about the observation actions? The executability precondition of $\mathit{observe}(A)$ being A, to make A executable amounts to making A true in the actual situation. Hence we have in this case the axiom

$$\mathit{After}_{\mathit{enable}_{\mathit{observe}(A)}}A \qquad (\text{TestAct}_6)$$

4.2 The axiom for revision

We are now able to postulate the following axiom for uninformative actions, which applies when revision is needed:

$$(\mathit{perc}(a,b) \wedge \mathit{BelAfter}_b\bot \wedge \neg\mathit{After}_a\bot) \rightarrow (\mathit{Feasible}_a\mathit{BelA} \leftrightarrow \mathit{BelAfter}_{\mathit{enable}_b}\mathit{After}_bA) \qquad (\text{SSA}_2)$$

Semantically, this means that when the agent is notified that b has occurred, and believes that b is inexecutable, then the possible situations after a are obtained by:

1. enabling b in every possible situation;
2. applying b to these situations;
3. collecting the resulting situations.

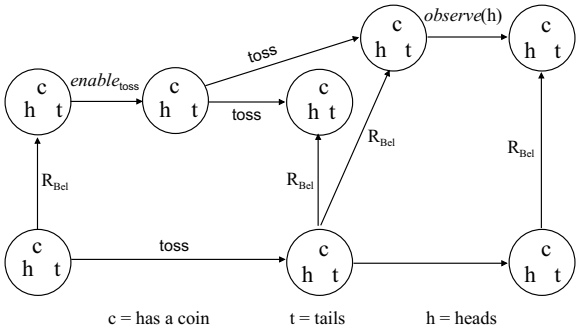

Figure 1. the toss example

Let us illustrate (SSA_2) by our running example. One of the possible Kripke models is given in figure 1. Suppose initially the agent ignores that there is a coin: $\mathit{HasCoin} \wedge \mathit{Bel}\neg\mathit{HasCoin}$. The tossing action is therefore executable, but the agent believes it isn't.

Suppose the coin is tossed resulting in Heads, and suppose the agent is correctly notified that toss has been executed: $\mathit{perc}(\mathsf{toss}, \mathsf{toss})$. In a first step the agent enables tossing by making its executability condition $\mathit{HasCoin}$ true, and then mentally executes toss. Putting these two actions together produces the resulting belief state, which is composed of a world where heads holds, and another one where tails holds. Syntactically, from (SSA_2) and the laws for toss we obtain $\mathit{After}_{\mathsf{toss}}\mathit{Bel}(\mathsf{Heads} \vee \mathsf{Tails})$, i.e. the agent believes the coin fell either heads or tails.

When the agent subsequently perceives that heads holds (via learning the occurrence of $\mathit{observe}(\mathsf{Heads})$) he then eliminates the world where tails holds from his belief state.

From SSA_1 and (SSA_2) one can derive a principle of *doxastic determinism*:

Theorem 2 $\mathit{Feasible}_a\mathit{BelA} \rightarrow \mathit{After}_a\mathit{BelA}$

5 Preserving facts

Given our successor state axioms we can reuse non-epistemic solutions to the frame problem.

Just as Scherl and Levesque have applied Reiter's solution [17] we use the solution of [3] in order to stay within propositional logic.

Which truths can be preserved after the performance of an uninformative action? Our key concept is that of the *influence of an action*. If there exists a relation of influence between the action and an atom p, then p cannot be preserved. The relation $a \rightsquigarrow p$ is read "the action a influences the truth value of p". In our example, $\rightsquigarrow = \{\mathsf{toss} \rightsquigarrow \mathsf{Heads}, \mathsf{toss} \rightsquigarrow \mathsf{Tails}, \mathsf{toss} \rightsquigarrow \mathit{HasCoin}\}$. Note that $\rightsquigarrow$ is in the metalanguage. We extend $\rightsquigarrow$ to formulas by stipulating that $a \rightsquigarrow A$ if there is an atom p occurring in A such that $a \rightsquigarrow p$.

The concept of influence (or dependence) is close to notions that have recently been studied in the field of reasoning about actions in order to solve the frame problem, e.g. Sandewall's [16] occlusion, Thielscher's [22] influence relation, or the 'possibly changes' operators of Giunchiglia *et al.* [7].

The preservation of formulas that are not influenced by an action is formalized by the influence-based *logical axiom*

$$A \rightarrow \mathit{After}_aA \text{ if } a \not\rightsquigarrow A \text{ and } A \text{ is factual} \qquad (\text{Preserv})$$

This expresses that if a does not influence A then A is preserved. The restriction that A be factual avoids e.g. $\mathit{Feasible}_{a'}\top \rightarrow \mathit{After}_a\mathit{Feasible}_{a'}\top$, which is not necessarily the case because a might modify the executability preconditions of a'.

6 Discussion and related work

We have defined a modal logic of belief and nondeterministic actions where the agent's beliefs about the action laws might be inaccurate. Our central axioms (SSA_1) and (SSA_2) have the form of successor state axioms. When actions are deterministic, (SSA_1) is exactly the syntactic counterpart of the successor state axiom of [17].

In our framework belief-contravening information can be restricted to learning that some action a has been executed. Inconsistency with the agent's beliefs means that the agent believes a to be inexecutable, and learns that a has occurred. We have shown that such a revision operation can be implemented by an updating operation enabling the execution of a. Our second axiom (SSA_2) is a new solution that does not resort to orderings of plausibility.

6.1 Regression.

When restricted to deterministic actions our axioms allow for regression. In the case of nondeterministic actions it is not clear how this could be done. An alternative is to use the famous modular completeness result due to Sahlqvist [15], which applies here almost immediately (because our axioms are of the required form). We thus

get for free soundness and completeness results, as well as a tableau algorithm. If the tableau algorithm terminates then we get a decision procedure for our logic. We are currently working on that, aiming at applying recent results on modal axioms of confluence and permutation (of which our SSA_1 and SSA_2 are instances).

6.2 Public actions

Almost all the approaches suppose that actions are public. It has been relaxed in [5], where drawbacks of the earlier solution in [12] are pointed out. The solution of [5] corresponds to our case where $perc(a, \lambda)$ holds.

6.3 Revision: the approach of Shapiro *et al.*

In [21], Shapiro *et col.* add to the Scherl and Levesque framework a revision-like operation based on plausibility orderings. They define $BelA$ as truth of A in the most plausible among the possible situations. If a sensing action eliminates the most plausible of the possible situations, then previously less plausible situations become the most plausible ones. The plausibility ordering should be kept fixed.

While being intuitively appealing, their solution has several drawbacks. (1) As the authors note, it is restricted to deterministic actions. (2) "The specification of [the plausibility ordering] over the initial situation is the responsibility of the axiomatizer of the domain." [21] This is particularly demanding because (3) in order to guarantee that after a the set of possible situations is nonempty, the authors require the set of possible situations to contain enough situations initially, restricting thus the agent's 'doxastic freedom'. (4) As pointed out in [5], such a solution to the problem of revision might endanger the solution to the frame problem. It seems to be fair to say that specifying a satisfactory plausibility ordering is a delicate task, involving a lot of imponderabilities in what concerns the relative plausibility of independent propositions. (5) The approach is unsatisfactory when applied to communication. Consider the following example: agent k is competent at p, and j is not. Agent i is completely ignorant initially: hence all possible situations are equally plausible for i. Then (under adequate hypotheses of cooperation) we can expect that when j asserts p, then i adopts p, i.e. $After_{asserts(j,p)}Belp$. Moreover, as all situations were equally plausible, p holds in every situation possible for i. Therefore when subsequently k asserts $\neg p$, i will unavoidably move to an empty set of possible situations. (6) Action occurrences are supposed to be perceived correctly and completely (and the agent is aware of that). Therefore wrong beliefs can only come from the initial situation, and the doxastic concept of [21] turns out to be quite close to knowledge.

6.4 Segerberg

The approach of Segerberg [19, 18] is similar in spirit to ours. He has a successor state axiom for expansion [19, axiom #12], but no such account for revision.

6.5 The AGM postulates

The normative framework for belief revision being the AGM theory [1], which of their postulates do we satisfy? With a similar encoding as that of Shapiro *et col.* it can be shown that we satisfy the basic postulates (K*1) – (K*4), and (K*6). (The names of the postulates are as in [21]). If we define update actions as in [21] we satisfy the update postulates (K$\diamond$1), (K$\diamond$2), (K$\diamond$4), and (K$\diamond$5) just as there. If we define updating by A as $enable_{observe(A)}$ then we moreover satisfy (K$\diamond$3).

REFERENCES

[1] Carlos E. Alchourrón, Peter Gärdenfors, and David Makinson, 'On the logic of theory change: Partial meet contraction and revision functions', *The Journal of Symbolic Logic*, **50**(2), (June 1985).

[2] F. Bacchus, J.Y. Halpern, and H. Levesque, 'Reasoning about noisy sensors in the situation calculus', in *Proc. 14th Int. Joint Conf. on Artificial Intelligence (IJCAI'95)*, pp. 1933–1940, (1995).

[3] Marcos A. Castilho, Olivier Gasquet, and Andreas Herzig, 'Formalizing action and change in modal logic I: the frame problem', *Journal of Logic and Computation*, **9**(5), 701–735, (1999).

[4] Silvia Coradeschi, 'Reasoning with misperception in the Features and Fluents Framework', in *Proc. 12th European Conf. on Artificial Intelligence (ECAI-96)*, pp. 657–661, (1996).

[5] Robert Demolombe and Maria del Pilar Pozos Parra, 'Formalisation de l'évolution de croyances dans le Calcul des Situations', in *Proc. of Modèles Formels de l'Interaction (MFI'01)*, eds., B. Chaib-draa and P. Enjalbert, volume 2, pp. 205–217, (2001).

[6] Gerbrandy, *Bisimulations on Planet Kripke*, Ph.D. dissertation, University of Amsterdam, 1999.

[7] E. Giunchiglia, G. N. Kartha, and V. Lifschitz, 'Representing action: indeterminacy and ramifi cations', *AI Journal*, **95**, (1997).

[8] David Harel, 'Dynamic logic', in *Handbook of Philosophical Logic*, eds., D. Gabbay and F. Guenthner, volume II, D. Reidel Publishing Company, (1984).

[9] Andreas Herzig, Jérôme Lang, and Thomas Polacsek, 'A modal logic for epistemic tests', in *Proc. Eur. Conf. on Artificial Intelligence (ECAI'2000)*, Berlin, (August 2000).

[10] Andreas Herzig, Jérôme Lang, Dominique Longin, and Thomas Polacsek, 'A logic for planning under partial observability', in *Proc. Seventeenth National Conf. on Artificial Intelligence (AAAI-2000)*, (2000).

[11] Hirofumi Katsuno and Alberto O. Mendelzon, 'Propositional knowledge base revision and minimal change', Technical Report KRR-TR-90-3, Dep. of Computer Science, University of Toronto, (March 1990).

[12] Yves Lespérance, Hector J. Levesque, and Raymond Reiter, 'A Situation Calculus approach to modeling and programming agents', in *Foundations and theories of Rational Agents*, eds., A. Rao and M. Wooldridge. Kluwer Academic Publishers, (1999).

[13] Hector J. Levesque, 'What is planning in the presence of sensing?', in *Proc. of the 13th Nat. Conf. on Artificial Intelligence (AAAI-96)*, Portland, Oregon, (August 1996).

[14] Robert C. Moore, 'A formal theory of knowledge and action', in *Formal Theories of the Commonsense World*, eds., J.R. Hobbs and R.C. Moore, 319–358, Ablex, Norwood, NJ, (1985).

[15] H. Sahlqvist, 'Completeness and correspondence in the fi rst and second order semantics for modal logics', in *Proc. 3rd Scandinavian Logic Symposium*, ed., S. Kanger, volume 82 of *Studies in Logic*, (1975).

[16] E. Sandewall, 'The range of applicability of some nonmonotonic logics for strict inertia', *J. of Logic and Computation*, **4**(5), 581–615, (1994).

[17] Richard Scherl and Hector J. Levesque, 'The frame problem and knowledge producing actions', in *Proc. Nat. Conf. on AI (AAAI'93)*, pp. 689–695. AAAI Press, (1993).

[18] Krister Segerberg, 'Two traditions in the logic of belief: bringing them together'. to appear.

[19] Krister Segerberg, 'Belief revision from the point of view of doxastic logic', *Bulletin of the IGPL*, **3**, 534–553, (1995).

[20] S. Shapiro, Yves Lespérance, and Hector J. Levesque, 'Specifying communicative multi-agent systems', in *Agents and Multi-Agent Systems - Formalisms, Methodologies, and Applications*, eds., W. Wobcke, M. Pagnucco, and C. Zhang, pp. 1–14. Springer-Verlag, LNAI 1441, (1998).

[21] S. Shapiro, M. Pagnucco, Y. Lespérance, and H. J. Levesque, 'Iterated belief change in the situation calculus', in *Proc. Seventh Int. Conf. on Principles of Knowledge Representation and Reasoning (KR 2000)*, pp. 527–538, (2000).

[22] Michael Thielscher, 'Computing ramifi cations by postprocessing', in *Proc. 14th Int. Joint Conf. on Artificial Intelligence (IJCAI'95)*, pp. 1994–2000, Montreal, Canada, (1995).

[23] Michael Thielscher, 'Representing the knowledge of a robot', in *Proc. KR'00*, eds., A. Cohn, F. Giunchiglia, and B. Selman, pp. 109–120. Morgan Kaufmann, (2000).

4.2 Social concepts

Electronic Notes in Theoretical Computer Science 157 (2006) 117–137

A Logical Framework for Grounding-based Dialogue Analysis[1]

Benoit Gaudou[a,2] Andreas Herzig[a,3] Dominique Longin[a,b,4]

[a] *Université Paul Sabatier*
IRIT – Équipe LILaC
118 route de Narbonne, 31062 Toulouse cedex 9, FRANCE

[b] *Université Toulouse-le-Mirail*
Laboratoire Travail et Cognition
Maison de la Recherche
5, allées Antonio Machado, 31058 Toulouse cedex 9, FRANCE

Abstract

A major critique against BDI (Belief, Desire, Intention) approaches to communication is that they require strong hypotheses such as sincerity and cooperation on the mental states of the agents (cf. for example [13,14,5]). The aim of this paper is to remed this defect. Thus we study communication between heterogeneous agents *via* the notion of *grounding*, in the sense of being publicly expressed and established. We show that this notion is different from social commitment, from the standard mental attitudes, and from different versions of common belief. Our notion is founded on speech act theory, and it is directly related to the *expression of the sincerity condition* [9,11,16] when a speech act is performed. We use this notion to characterize speech acts in terms of preconditions and effects. As an example we show how persuasion dialogues *à la Walton & Krabbe* can be analyzed in our framework. In particular we show how speech act preconditions constrain the possible sequences of speech acts.

Keywords: grounding, commitment, dialogue, speech acts, modal logic, common belief, BDI logic

[1] Thanks to the anonymous referees of LCMAS'05 who have helped to increase the quality of this paper. Thanks to Nicolas Maudet and Philippe Muller for useful discussions.
[2] Email: Benoit.Gaudou@irit.fr
[3] Email: Andreas.Herzig@irit.fr
[4] Email: Dominique.Longin@irit.fr

1 Introduction

Traditionally there are two ways to analyze dialogues: the first one is through their structure, and the second one is through the participants' mental states. The former approaches analyze dialogues independently of the agents' mental states and focus on what a third party would perceive of it. This route is taken by the *conventional* approaches such as Conte and Castelfranchi's [3], Walton and Krabbe's dialogue games [19], Singh [14], and Colombetti et col. [5,18], who study the notion of social commitment.

On the one hand, a major critique concerning the mentalist approaches (cf. e.g. [14,5]) is that they require strong hypotheses on the architecture of the agents' internal state and the principles governing their behavior (such as sincerity, cooperation, competence), while agents communicating in open systems are heterogeneous and might thus work with very different kinds of internal states and principles. Suppose for example a speaker asserts that p. Then he may or may not believe that p, depending on his sincerity. The hearer may or may not believe the speaker believes p, depending his beliefs about the speaker's sincerity. And if the hearer starts to believe the speaker believes p, the hearer may or may not start to believe p himself depending on his beliefs about the speaker's competence.

On the other hand, a common hypothesis in formal frameworks for agent-to-agent communication is to suppose speech acts are public, and that there is no misperception in dialogue: perception of speech acts is sound and complete with respect to reality.

In this paper we propose a *notion of grounding* which captures what is expressed and established during a conversation between different agents (Sect. 2). Using a particular modal operator to capture this notion (Sect. 3), we show that it is at the borderline between mentalist and structure-based approaches (Sect. 4). We then study a particular kind of dialogue (Walton and Krabbe's PPD_0 persuasion dialogues) by characterizing the speech act types that are involved there (Sect. 5). Our characterization induces a protocol governing the conversational moves. Contrarily to what is usually done in Agent Communication Languages (ACL) this protocol is not described in some metalanguage but on the object language level.

2 Grounding

Here we investigate the notion of grounded information, which we view as information that is *publicly expressed and accepted as being true by all the agents participating in a conversation*. A piece of information might be grounded

even when some agents privately disagree, as long as they do not manifest their disagreement.

Our notion stems from speech act theory, where Searle's *expression of an Intentional state* [11] is about a psychological state related to the state of the world. Even if an utterance was unsincere an Intentional State has been expressed, and that state corresponds in some way to a particular belief of the speaker.

The concept of groundedness applies to Moore's paradox, according to which one cannot successfully assert "p is true and I do not believe p". The paradox follows from the fact that: on the one hand, the assertion entails expression of the sincerity condition about p (the speaker believes p); on the other hand, the assertion expresses the speaker believes he believes p is false. Via a principle of introspection this expresses that he believes p is false, and the assertion is contradictory.

Vanderveken [16,17] has captured the subtle difference between *expressing* an Intentional state and *really being in* such a state by distinguishing *success conditions* from *non-defectiveness conditions*, thus refining the felicity conditions as defined by Searle [9,10,12]. According to Vanderveken, when we assert p we *express* that we believe p (success condition), while the speaker's being in a state of believing that p is a condition of non-defectiveness.

Whenever an agent asserts p then it is grounded that he believes that p, independently of the agent's individual beliefs. For a group of agents we say that a piece of information is grounded if and only if for every agent it is grounded that he believes it.

Groundedness is an objective notion: it refers to what can be observed, and only to that. While it is related to mental states because it corresponds to the expression of Intentional states, it is not an Intentional state: it is neither a belief nor a goal, nor an intention. As we shall see, it is simple and elegant way of characterizing mutual belief.

We believe that such a notion is interesting because it fits the public character of speech act performance. As far as we are aware the logical investigation of such a notion has neither been undertaken in the social approaches nor in the conventional approaches. A similar notion has been investigated very recently in [7], which formalizes the notions of manifested opinion in the sense of ostensible belief and of ostensible intention.

3 Logical framework

In this section, we present a light version of the logic of belief, choice and intention we developed in [6], augmented by a modal operator expressing

"groundedness". In particular, we neither develop here temporal aspects nor relations between action and mental attitudes (the frame problem for belief and choice).

3.1 Semantics

Let $AGT = \{i, j, \ldots\}$ be a set of agents. We suppose AGT is finite. Let $ATM = \{p, q, \ldots\}$ be the set of propositions. Complex formulas are denoted by $A, B, C, \ldots$. A model includes a set of possible worlds W and a mapping $V : W \rightarrow (ATM \rightarrow \{0, 1\})$ associating a valuation V_w to every $w \in W$. Models moreover contain accessibility relations that will be detailed in the sequel.

Belief.

In order to not only speak about facts, but also the participants' beliefs we introduce a modal operator of belief. $Bel_i A$ reads "agent i believes that A holds", or "agent i believes A". To each agent i and each possible world w we associate a set of possible worlds $\mathcal{B}_i(w)$: the worlds that are consistent with i's beliefs. The function $\mathcal{B}_i$ can be viewed as an accessibility relation. As usual the truth condition for Bel_i stipulates that it holds that A is believed by agent i at w, noted $w \Vdash Bel_i A$, iff A holds in every $w' \in \mathcal{B}_i(w)$. We suppose that:

❶ $\mathcal{B}_i$ is serial, transitive and euclidian.

Grounding.

GA reads "it is grounded (for the considered group of agents) that A is true" (or for short : "A is grounded"). Grounded here means public and agreed by everybody. To each world w we associate the set of possible worlds $\mathcal{G}(w)$ that are consistent with all grounded propositions. $\mathcal{G}(w)$ contains those worlds where all grounded propositions hold. The truth condition for G stipulates that it holds that A is grounded in w, noted $w \Vdash GA$, iff A holds in every $w' \in \mathcal{G}(w)$. Just as the $\mathcal{B}_i$, $\mathcal{G}$ can be viewed as an accessibility relation. We suppose that

❷ $\mathcal{G}$ is serial, transitive and euclidian.

Belief and grounding.

We postulate the following relationship between the accessibility relations for $\mathcal{B}_i$ and G:

❸ if $w' \in \mathcal{B}_i(w)$ then $\mathcal{G}(w) = \mathcal{G}(w')$

❹ if $u\mathcal{G}v$ and $v\mathcal{B}_i w$ then there is w' such that $w\mathcal{G}w'$ and $V(w) = V(w')$

❺ $\mathcal{G} \subseteq \mathcal{G} \circ \bigcup_{i \in AGT} \mathcal{B}_i$

The constraint ❸ stipulates that agents are aware of what is grounded and of what is ungrounded.

The constraint ❹ stipulates that for every grounded proposition it is publicly established that every agent believes it (which does not imply that they actually believe them): whenever w is a world for which all believed propositions of agent i are grounded, then all those propositions are indeed grounded in w.

The constraint ❺ expresses that if a proposition is established for every agent (*i.e.* it is grounded that every agent believes it) then it is grounded: whenever w is a world for which all grounded propositions hold, then it is indeed grounded that it is possible, for every agent, that all these propositions hold in w.

Choice.

Among all the worlds in $\mathcal{B}_i(w)$ that are possible for agent i, there are some that i prefers. Cohen and Levesque [2] say that i *chooses* some subset of $\mathcal{B}_i(w)$. Semantically, these worlds are identified by yet another accessibility relation

$$\mathcal{C}_i : W \to 2^W$$

$Ch_i A$ expresses that agent i chooses that A. We sometimes also say that i *prefers* that A[5]. Without surprise, $w \models Ch_i A$ if A holds in all preferred worlds, i.e. $w \models Ch_i A$ if $w' \models A$ for every $w' \in \mathcal{C}_i(w)$. We suppose that

❻ $\mathcal{C}_i$ is serial, transitive, and euclidian.[6]

Choice and belief, choice and grounding.

As said above, an agent only chooses worlds he considers possible:

❼ $\mathcal{C}_i(w) \subseteq \mathcal{B}_i(w)$.

Hence belief implies choice, and choice is a mental attitude that is weaker than belief. This corresponds to validity of the principle $Bel_i A \to Ch_i A$.

We moreover require that worlds chosen by i are also chosen from i's possible worlds, and *vice versa* (see Figure 1):

[5] While Cohen and Levesque use a modal operator 'goal' (probably in order to have a uniform denomination w.r.t. the different versions of goals they study), it seems more appropriate to us to use the term 'choice'.

[6] This differs from Cohen and Levesque, who only have supposed seriality, and follows Sadek's approach. The latter [8] has argued that choice is a mental attitude which obeys to principles of introspection that correspond with transitivity and euclideanity.

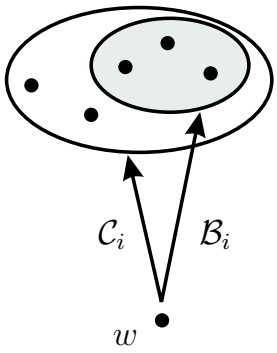

Fig. 1. Belief and choice

❽ if $w\mathcal{B}_i w'$ then $\mathcal{C}_i(w) = \mathcal{C}_i(w')$.

We do not suppose any semantical constraint between choice and grounding beyond those coming with the above $\mathcal{C}_i(w) \subseteq \mathcal{B}_i(w)$.

Action.

Let $ACT = \{\alpha, \beta \ldots\}$ be the set of actions. Speech acts are particular actions; they are 4-uples of the form $\langle i, j, FORCE, A\rangle$ where i is the author of the speech act, j its addressee, $FORCE$ its illocutionary force, and A a formula denoting its propositional content. For example $\langle i, j, \mathsf{Assert}, p\rangle$ expresses that i asserts to j that p is true. We write α_i to denote that i is the author of α.

The formula $After_\alpha A$ expresses that if α happens then A holds after α. The dual $Happens_\alpha A = \neg After_\alpha \neg A$ means that α happens and A is true afterwards. Hence $After_\alpha \bot$ expresses that α does not happen, and $Happens_\alpha \top$ that α happens and we write then $Happens(\alpha)$. For every action $\alpha \in ACT$ there is a relation $R : ACT \to (W \to 2^W)$ associating sets of worlds $R_\alpha(w)$ to w. The truth condition is: $w \Vdash After_\alpha A$ iff $w' \Vdash A$ for every $w' \in R_\alpha(w)$.

The formula $Before_\alpha A$ means that before every execution of α, A holds. The dual $Done_\alpha A = \neg Before_\alpha \neg A$ expresses that the action α has been performed before which A held. Hence $Done_\alpha \top$ means that α just has happened. The accessibility relation for $Before_\alpha$ is the converse of the above relation R_α. The truth condition is: $w \Vdash Before_\alpha A$ iff $w' \Vdash A$ for every $w' \in R_\alpha^{-1}(w)$.

As said above, we do not detail here the relationship between action and mental attitudes and refer the reader to [6].

Belief.

The axioms corresponding to the semantical conditions for belief are those of KD45, *i.e.* those of normal modal logics [1], plus the following:

$$Bel_i A \to \neg Bel_i \neg A \qquad (\mathrm{D}_{Bel_i})$$
$$Bel_i A \to Bel_i Bel_i A \qquad (4_{Bel_i})$$
$$\neg Bel_i A \to Bel_i \neg Bel_i A \qquad (5_{Bel_i})$$

Hence an agent's beliefs are consistent (D_{Bel_i}), and he is aware of his beliefs (4_{Bel_i}) and disbeliefs (5_{Bel_i}). The following are theorems of the logic:

$$Bel_i A \leftrightarrow Bel_i Bel_i A \qquad (1)$$
$$Bel_i \neg Bel_i A \leftrightarrow \neg Bel_i A \qquad (2)$$

Grounding.

The logic of the grounding operator is again a normal modal logic of type KD45:

$$GA \to \neg G \neg A \qquad (\mathrm{D}_G)$$
$$GA \to GGA \qquad (4_G)$$
$$\neg GA \to G \neg GA \qquad (5_G)$$

(D_G) expresses that the set of grounded informations is consistent: it cannot be the case that both A and $\neg A$ are simultaneously grounded.

(4_G) and (5_G) account for the public character of G. From these *collective awareness* results: if A has (resp. has not) been grounded then it is established that A has (resp. has not) been grounded.

The following theorems follow from (D_G), (4_G), and (5_G):

$$GA \leftrightarrow GGA \qquad (3)$$
$$G \neg GA \leftrightarrow \neg GA \qquad (4)$$

Belief and grounding.

In accordance with the preceding semantic conditions the following are logical axioms:

$$GA \to Bel_i GA \qquad (\text{SR}_+)$$
$$\neg GA \to Bel_i \neg GA \qquad (\text{SR}_-)$$
$$G\varphi \to GBel_i\varphi, \text{ for } \varphi \text{ factual} \qquad (\text{WR})$$
$$(\bigwedge_{i \in AGT} GBel_i A) \to GA \qquad (\text{CG})$$

where a factual formula does not contain any modality.

(SR_+) and (SR_-) together correspond to ❸. (WR) corresponds to ❹, and (CG) to ❺.

The axioms of strong rationality (SR_+) and (SR_-) express that the agents are aware of the grounded (resp. ungrounded) propositions (cf. (5) and (6) below). This is due to the public character of the grounding operator.

(WR) expresses that if the factual formula φ is grounded then it is necessarily grounded that each agent expressed that he believes φ[7]. Note that this does not imply that every agent actually believe it, *i.e.* (WR) does not entail $G\varphi \to Bel_i\varphi$.

(WR) concerns only factual formulas. When an agent performs the speech act $\langle i, j, \mathsf{Assert}, p\rangle$, he expresses publicly that he believes p. ($Bel_i p$ is publicly established so $GBel_i p$ holds.) This does not mean that i indeed believes p: i might ignore whether p, or even believe that $\neg p$. It would be hypocritical to impose that it is grounded for another agent j that $Bel_i p$. Therefore $GBel_i p \to GBel_j Bel_i p$ should not be valid. Moreover, if we applied (WR) to some mental states, we would restrict the agents' autonomy. For example, when agent i performs the speech act: $\langle i, j, \mathsf{Assert}, Bel_j p\rangle$ then afterwards the formula $GBel_i Bel_j p$ holds, and the agent j could not later on express that he believes $\neg p$. Indeed, if he made this speech act, the formulae $GBel_j \neg p$ and, thanks to (WR), $GBel_i Bel_j \neg p$ would hold, which is inconsistent with the above formula $GBel_i Bel_j p$.

(CG) expresses that if a proposition is established for every agent in AGT then it is grounded.

[7] This axiom does not presuppose that an agent i explicitly asserted φ, even if, in our current theory, we do not describe the mechanism of an agent's implicit commitment. Moreover, for Walton & Krabbe [19], agents can not incur implicitly strong commitments. (We will show in Section 5 links between grounding, belief and commitments *à la* Walton & Krabbe.)

The followings are straightforward consequences of (SR_+) and (SR_-):

$$GA \leftrightarrow Bel_i GA \tag{5}$$
$$\neg GA \leftrightarrow Bel_i \neg GA \tag{6}$$

These theorems express that agents are aware of what is grounded.

Choice.

Similar to belief, we have the (D_{Ch_i}), (4_{Ch_i}) and (5_{Ch_i}). (See [6] for more details.)

Choice and belief.

Our semantics validates the equivalences:

$$Ch_i A \leftrightarrow Bel_i Ch_i A \tag{7}$$
$$\neg Ch_i A \leftrightarrow Bel_i \neg Ch_i A \tag{8}$$

This expresses that agents are aware of their choices.

Action.

As the relation $R_\alpha^{-1}(w)$ is the converse of R_α, we have the two conversion axioms:

$$A \rightarrow After_\alpha Done_\alpha A \qquad (\mathrm{I}_{After_\alpha, Done_\alpha})$$
$$A \rightarrow Before_\alpha Happens_\alpha A \qquad (\mathrm{I}_{Before_\alpha, Happens_\alpha})$$

3.3 Action laws

Action laws for an action come in two kinds: *executability laws* describe the preconditions of the action, and *effect laws* describe the effects. The preconditions of an action are the conditions that must be fulfilled in order that the action is executable. The effects (or postconditions) are properties that hold after the action because of it. For example, to toss a coin, we need a coin (precondition) and after the toss action the coin is heads or tails (postcondition).

The set of all action laws is noted $LAWS$, and some examples are collected in Table 2. The general form of an executability law is:

$$Ch_i Happens(\alpha_i) \wedge precond(\alpha_i) \leftrightarrow Happens(\alpha_i) \qquad (\mathrm{Int}_{Ch_i, \alpha_i})$$

This expresses a principle of intentional action: an action happens exactly when its preconditions hold and its author chooses it to happen. The general form of an effect law is $A \rightarrow After_{\alpha} postcond(\alpha)$. In order to simplify our exposition we suppose that effect laws are unconditional and therefore the general form of an effect law is here:

$$After_{\alpha} postcond(\alpha)$$

A way of capturing the conventional aspect of interaction is to suppose that these laws are common to all the agents. Formally they are thus global axioms to which the necessitation rule applies [4].

4 Groundedness compared to other notions

In our formalism, $GBel_i A \rightarrow Bel_i A$ is not valid. Thus, when it is grounded that a piece of information A holds for agent i then this does not mean that i indeed believes that A. The other way round, $Bel_i A \rightarrow GBel_i A$ is not valid either: an agent might believe A while it is not grounded that A holds for i.

The operator G is objective in nature. It is different from other objective operators such as that of social commitment of [13,14,5,18]. To see this consider speech act semantics: as we have shown (cf. Sect. 2), the formula $GBel_i A$ expresses the idea that it is grounded that A holds for agent i. This has to be linked to the expression of an Intentional state as a necessary condition for the performance of a speech act. This means that when agent i asks agent j to pass him the salt then it has been established either that i wants to know whether j is able to pass him the salt (literal meaning), or that i wants j to pass him the salt (indirect meaning). In a commitment-based approach this typically leads to a conditional commitment (or precommitment) of j to pass the salt, which becomes an unconditional commitment upon a positive reaction. In our approach we do not try to determine whether j *must* do such or such action: we just establish the facts, without any hypothesis on the agents' beliefs, goals, intentions, ... or commitments.

On the other hand, as the next section shows, some obligations that can be found in commitment-based approaches have a counterpart in our formalism: our characterization of speech acts in terms of preconditions and effects constrains the agents' options for the choice of actions, as well as their order (cf. Sect. 5).

In fact, the operator G expresses a sort of common belief. In [15], Tuomela distinguishes *(proper) group beliefs* from *shared we-beliefs*. In the first case a group may typically believe a proposition while none of the agents of the group really believes it. In the second case, the group holds a belief which

each individual agent really holds, too.

Our operator G is closer to Tuomela's *(proper) group beliefs* because the formula $GA \rightarrow Bel_i A$ is invalid. Thus, GA means that a group [Agt] "(intentionally) jointly accept A as the view of [Agt] (...) and there is a mutual belief [about this]" [15]. Different from Tuomela we do not distinguish the agents contributing to the grounding of the group belief from those which passively accept it.

5 Walton&Krabbe's persuasion dialogues (PPD_0)

We now apply our formalism to a particular kind of dialogue, viz. persuasion dialogues. We characterize the speech acts of Walton&Krabbe's (W&K for short) game of dialogue PPD_0, also called Permissive Persuasion Dialogue. These works mainly follow from Hamblin's works. In order to simplify our exposition we suppose that there are only two agents (but the account can easily be generalized to n agents).

A persuasion dialogue takes place when there is a conflict between two agents' beleif. The goal of the dialogue is to resolve this situation: an agent can persuade the other party to concede his own thesis (in this case he wins the dialogue game) or concede the point of view of the other party (and thus lose the game).

W&K distinguish two kinds of commitment: those which can be challenged (*assertions*) and those which cannot (*concessions*). We formalize this distinction with the notions of strong commitment (SC) and weak commitment (WC). They are linked by the fact that a strong commitment to a proposition implies a weak commitment to it ([19, p. 133]). We use the logical framework presented above to formalize these two notions, and apply it to PPD_0. In relation with this logical framework, we define: [8]

$$SC_i\, A \stackrel{def}{=} GBel_i A \qquad (\mathrm{Def}_{SC_i})$$

$$WC_i\, A \stackrel{def}{=} G\neg Bel_i \neg A \qquad (\mathrm{Def}_{WC_i})$$

Note that we might have chosen to have primitive operators SC_i, and define GA as being an abbreviation of $(\bigwedge_{i \in AGT} SC_i\, A)$.

In terms of the preceding abbreviations we can prove:

[8] This is an approximation of W&K's *assertion*. Indeed, our $GBel_i A$ is "more logical" than W&K's *a(A)*: W&K allow both $a(A)$ and $a(\neg A)$ to be the case simultaneously, while for us $GBel_i A \wedge GBel_i \neg A$ is inconsistent. In the case of weak commitment, we agree with W&K's works: in our framework, $WC_i\, A \wedge WC_i\, \neg A$ is consistant.

$$SC_i\, A \rightarrow \neg SC_i \neg A \tag{9}$$
$$SC_i\, A \leftrightarrow SC_i\, SC_i\, A \tag{10}$$
$$\neg SC_i\, A \leftrightarrow SC_i \neg SC_i\, A \tag{11}$$

(9) shows the rationality of the agents: they cannot commit both on A and $\neg A$. (10) and (11) account for the public character of commitment. With those three theorems, we can show that SC_i is an operator of a normal modal logic of type KD45, too. [9]

$$GA \leftrightarrow SC_i\, GA \tag{12}$$
$$\neg GA \leftrightarrow SC_i \neg GA \tag{13}$$
$$SC_i\, A \leftrightarrow SC_j\, SC_i\, A \tag{14}$$
$$\neg SC_i\, A \leftrightarrow SC_j \neg SC_i\, A \tag{15}$$

These theorems are some consequences of the public character of the commitment. (12) and (13) entail that it is grounded that the agents are committed to the grounded (resp. ungrounded) propositions. (14) and (15) mean that each agent is committed to the other agents commitments, and non-commitments.

$$SC_i\, A \rightarrow WC_i\, A \tag{16}$$
$$WC_i\, A \rightarrow \neg SC_i \neg A \tag{17}$$

(16) says that strong commitment implies weak commitment. (17) expresses that if agent i is weakly committed to A then i is not strongly committed to $\neg A$.

$$WC_i\, A \leftrightarrow SC_j\, WC_i\, A \tag{18}$$
$$\neg\, WC_i\, A \leftrightarrow SC_j \neg\, WC_i\, A \tag{19}$$

(18) expresses that weak commitment is public. (19) is similar for absence of weak commitment.

[9] We can prove that K is a theorem for SC_i and that the necessitation rule can be applied to it.

Precond(α)	Act α	Postcond(α)
$\neg SC_s\, p$	$\langle s, h, \mathsf{Assert}, p\rangle$	$SC_s\, p$
$SC_s\, p$	$\langle s, h, \mathsf{SRetract}, p\rangle$	$\neg SC_s\, p$
$WC_s\, p$	$\langle s, h, \mathsf{WRetract}, p\rangle$	$\neg WC_s\, p$
$SC_s\, p \wedge \neg WC_h\, p$	$\langle s, h, \mathsf{Argue}, (q_1, ..., q_n SOp)\rangle$	$\bigwedge_{1\leq i\leq n} SC_s\, q_i \wedge$ $SC_s\, (\bigwedge_{1\leq i\leq n} q_i \rightarrow p)$
$\neg WC_s\, p$	$\langle s, h, \mathsf{Concede}, p\rangle$	$WC_s\, p$
$\neg WC_s\, p$	$\langle s, h, \mathsf{RefuseConcede}, q\rangle$	$\neg WC_s\, p$
$SC_s\, q \wedge \neg WC_h\, q \wedge \neg WC_h\, p$	$\langle s, h, \mathsf{RequestConcede}, p\rangle$	$\emptyset$
$\neg WC_s\, p \wedge SC_h\, p \wedge$ $\neg GDone_{\langle s,h,\mathsf{Challenge},p\rangle} \top$	$\langle s, h, \mathsf{Challenge}, p\rangle$	$\emptyset$
$\neg WC_h\, p$	$\langle s, h, \mathsf{Serious}, p\rangle$	$\emptyset$
$WC_h\, p \wedge WC_h\, q \wedge (p \leftrightarrow \neg q)$	$\langle s, h, \mathsf{Resolve}, p\rangle$	$\emptyset$

Table 1
Preconditions and effects of speech acts (with commitments).

5.1 Speech acts and grounding

The dialogues that we want to formalize (W&K-like dialogues) are controlled by some conventions: the rules of the game. The allowed sequences of acts are those of W&K's PPD_0 (cf. [19, p. 150-151]). They are formalized in Figure 2 and will be discussed below. For example, after a speech act $\langle s, h, \mathsf{Assert}, p\rangle$,

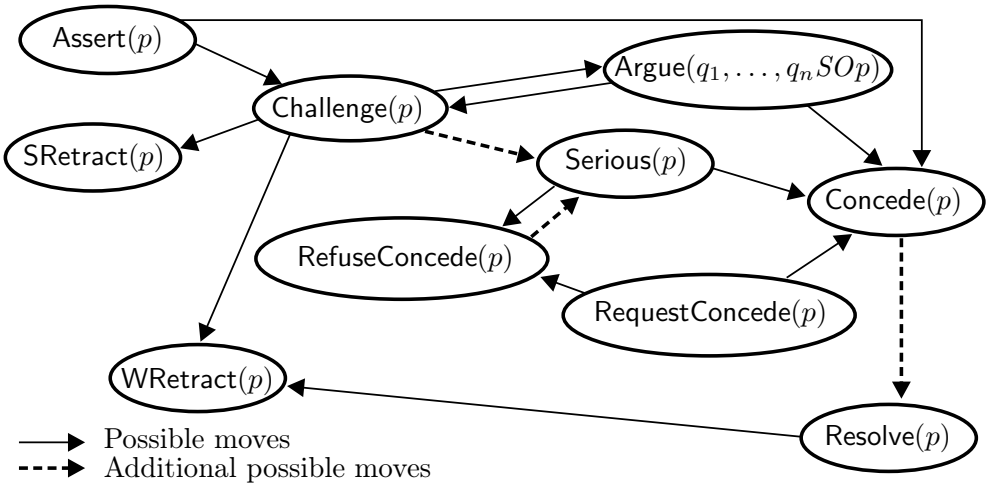

Fig. 2. (Additional) possible moves after each act

the hearer can only challenge p or concede it. We formalize them in our logic by expressing that an act grounds that the hearer's choices are limited only to some acts. Speech acts have two different effects: one is on the commitment store in terms of weak and strong commitments (cf. Table 1) and the other one is the set of acts the hearer can perform in response (cf. Table 2).

We suppose that initially nothing is grounded, *i.e.* the belief base is $\{\neg GA : A \text{ is a formula }\}$. [10]

Acts α	Constraints on the possible actions following α
$\langle s, h, \mathsf{Assert}, p\rangle$	$G(Ch_h Happens(\langle h, s, \mathsf{Challenge}, p\rangle) \vee$ $Ch_h Happens(\langle h, s, \mathsf{Concede}, p\rangle))$
$\langle s, h, \mathsf{SRetract}, p\rangle$	$\emptyset$
$\langle s, h, \mathsf{WRetract}, p\rangle$	$\emptyset$
$\langle s, h, \mathsf{RequestConcede}, p\rangle$	$G(Ch_h Happens(\langle h, s, \mathsf{RefuseConcede}, p\rangle) \vee$ $Ch_h Happens(\langle h, s, \mathsf{Concede}, p\rangle))$
$\langle s, h, \mathsf{Argue}, (q_1, ..., q_n SOp)\rangle$	$\bigwedge_{1 \leq i \leq n} G(Ch_h Happens(\langle h, s, \mathsf{Challenge}, q_i\rangle) \vee$ $Ch_h Happens(\langle h, s, \mathsf{Concede}, q_i\rangle))$ $\wedge G(Ch_h Happens(\langle h, s, \mathsf{Challenge}, q_1 \wedge ... \wedge q_n \rightarrow p\rangle) \vee$ $Happens(\langle h, s, \mathsf{Concede}, q_1 \wedge ... \wedge q_n \rightarrow p\rangle))$
$\langle s, h, \mathsf{Challenge}, p\rangle$	$G(Ch_h Happens(\langle h, s, \mathsf{SRetract}, p\rangle) \vee$ $Ch_h Happens(\langle h, s, \mathsf{WRetract}, p\rangle) \vee$ $Ch_h Happens(\langle h, s, \mathsf{Argue}, (q_1, ..., q_n SOp)\rangle) \vee$ $Ch_h Happens(\langle h, s, \mathsf{Serious}, p\rangle))$
$\langle s, h, \mathsf{Concede}, p\rangle$	$\emptyset$
$\langle s, h, \mathsf{RefuseConcede}, p\rangle$	$\emptyset$
$\langle s, h, \mathsf{Serious}, p\rangle$	$G(Ch_h Happens(\langle h, s, \mathsf{RefuseConcede}, p\rangle) \vee$ $Ch_h Happens(\langle h, s, \mathsf{Concede}, p\rangle))$
$\langle s, h, \mathsf{Resolve}, p\rangle$	$G(Ch_h Happens(\langle h, s, \mathsf{WRetract}, p\rangle) \vee$ $Ch_h Happens(\langle h, s, \mathsf{WRetract}, \neg p\rangle))$

Table 2
Additional postconditions of speech acts.

The Assert act on p can only be used by the two parties in some preliminary moves of the dialogue to state the theses of each participant. The effect of the act is that it is grounded that its content p holds for the speaker: he has expressed a kind of strong commitment (an *assertion* for W&K) on p in the sense that he must defend his commitment by an argument if it is challenged.

To Concede p means to admit that p could hold, where p has been asserted by the other party. The effect of this act is that it is grounded that the speaker has taken a kind of commitment on p. But the nature of this commitment is not the same as the former one: this one has not to be defended when it is attacked. W&K call it *concession* and it corresponds to our notion of Weak Commitment.

The Challenge act on p forces the other participant to either put forward an argument for p, or to retract the assertion p. For a given propositional

[10] This is an infinite set. In practice one would resort to default reasoning here.

content this act can only be performed once.

Argue: to defend a challenged assertion p, an argument must have p as conclusion and a set of propositions $q_1 ... q_n$ as premises. We write it as follows:

$$q_1 ... q_n SOp \stackrel{def}{=} q_1 \wedge ... \wedge q_n \wedge (q_1 \wedge ... \wedge q_n \rightarrow p) \qquad (\text{Def}_{SO})$$

The effect of this act is that all premises q_1, ..., q_n and the implicit implication $q_1 \wedge ... \wedge q_n \rightarrow p$ are grounded for the speaker. It follows that the challenger must explicitly take position in the next move (challenge or concede) on each premise and on the implicit implication. To challenge one premise means that the argument cannot be applied, while to challenge the implicit implication means that the argument is incorrect. If he does not challenge a proposition, he (implicitly) concedes it. But as soon as he has conceded all the premises and the implication, he must also concede the conclusion. To avoid some digressions, W&K suppose that an unchallenged assertion cannot be defended by an argument. Moreover, we took over their form of the support of arguments, viz. $A \rightarrow B$, although we are aware that more complex forms of reasoning occur in real world argumentation.

At any time, the speaker may request more concessions (with a Request-Concede act) from the hearer, to use them as premises for arguments. The hearer can then accept or refuse to concede.

W&K use the same speech act type to retract a concession and to refuse to concede something (the act $nc(p)$). But it seems to us that it is not the same kind of act, and we decided to create two different acts: $\langle s, h, \mathsf{WRetract}, p\rangle$ to retract one of his own weak commitments, and $\langle s, h, \mathsf{RefuseConcede}, p\rangle$ to decide not to concede anything. A strong commitment can be retracted with a $\langle s, h, \mathsf{SRetract}, p\rangle$. This act removes the strong commitment from the commitment store, but not the weak commitment, whereas the $\langle s, h, \mathsf{WRetract}, p\rangle$ act removes the weak commitment and, if it exists, the strong commitment, too.

In our logic, $WC_i\, A \wedge WC_i \neg A$ is satisfiable, but not $SC_i\, A \wedge SC_i \neg A$. Thus we are more restrictive than W&K: in the following, a contradiction in an agents' commitment store is only due to contradictory Weak Commitments. [11] When a party detects an contradiction in the other party's commitment store, it can ask him to resolve it (with the act Resolve(p,q) where "p and q are explicit contradictories" [19, p. 151].). The other party must retract one of the inconsistent propositions. W&K do not make any inference in the commitment

[11] W&K allow the agents to have some contradictory concessions (WC) and assertions (SC) in their commitment store (*i.e.* $SC_i\, A$ and $SC_i \neg A$ or $WC_i\, A$ and $WC_i \neg A$ can hold simultaneously).

store, so Resolve only applies to explicit inconsistency (that is: Resolve(p,$\neg p$)). We will write Resolve(p) instead of Resolve(p,q) where q is $\neg p$. (Resolve(p) and Resolve($\neg p$) are thus equivalent.) To perform the speech act Resolve(p), we can show that it is necessary and sufficient that the propositions p and $\neg p$ are weak commitments of the agent. In our formalism, the act Resolve holds only to weak commitments. Moreover the two contradictory weak commitments cannot be derived from two inconsistent strong commitments (which W&K allow), because such are consistent in our logic.

When an agent chooses to challenge a proposition p or to refuse to concede it, his opponent can query him to reassess his position. Finally the speech act Serious(p) imposes that the agent must concede p or refuse to concede it.

Note that W&K define another commitment store that contains what they call *dark-side commitments*. If p is a dark-side commitment, it must be revealed after a Serious(p) and the agent must concede p and cannot retract it. We do not consider such commitments here because, we focus on what is observable and objective in the dialogue: so if an agent chooses to concede p, we do not know if it was a dark-side commitment or not, consequently the agent may, even if it had a dark-side commitment on p and contrary to W&K's theory, retract it in a subsequent dialogue move dialogue.

The action preconditions are not mutually exclusive. This gives the agents some freedom of choice. We do not describe here the subjective cognitive processes that lead an agent to a particular choice.

5.2 Example

We recast an example of a persuasion dialogue given by W&K [19, p. 153] to illustrate the dialogue game PPD_0 (see Figure 3): initially, agent i asserts p_1 and agent j asserts p_2. Thus, the following preparatory moves have been performed: $\langle i, j, \mathsf{Assert}, p_1\rangle$ and $\langle j, i, \mathsf{Assert}, p_2\rangle$.

After each move, the agents' commitment stores are updated (see Table 3). In his first move, j asks i to concede p_3 and challenges p_1. i responds by conceding p_3, etc. In move (vii), agent j concedes p_1 which is the thesis of his opponent. [12]

As we have said, in order to stay consistent with our logical framework, we have to add an effect to the W&K speech act of concession: when i concedes a proposition p, every strong commitment of i on $\neg p$ is retracted. Agent i is then weakly committed on both p and $\neg p$. We thus weaken the paraconsistent aspects of W&K, viz. that an agent can have assertions or concessions that

[12] He thus loses the game in what concerns the thesis of i but in what concerns his own thesis, the game is not over yet.

(i) $\langle j, i, \mathsf{RequestConcede}, p_3\rangle$,
$\langle j, i, \mathsf{Challenge}, p_1\rangle$

(ii) $\langle i, j, \mathsf{Concede}, p_3\rangle$,
$\langle i, j, \mathsf{Serious}, p_1\rangle$,
$\langle i, j, \mathsf{Argue}, (p_3 SO p_1)\rangle$,
$\langle i, j, \mathsf{Challenge}, p_2\rangle$

(iii) $\langle j, i, \mathsf{RefuseConcede}, p_1\rangle$,
$\langle j, i, \mathsf{Concede}, p_3 \to p_1\rangle$,
$\langle j, i, \mathsf{Argue}, (p_4, p_5 SO p_2)\rangle$,
$\langle j, i, \mathsf{Challenge}, p_3\rangle$

(iv) $\langle i, j, \mathsf{Concede}, p_5\rangle$,
$\langle i, j, \mathsf{Concede}, p_4 \wedge p_5 \to p_2\rangle$,
$\langle i, j, \mathsf{Serious}, p_3\rangle$,
$\langle i, j, \mathsf{Argue}, (\neg p_4, p_5 SO p_3)\rangle$,
$\langle i, j, \mathsf{Challenge}, p_4\rangle$

(v) $\langle j, i, \mathsf{WRetract}, p_3 \to p_1\rangle$,
$\langle j, i, \mathsf{Concede}, p_3\rangle$,
$\langle j, i, \mathsf{Concede}, \neg p_4\rangle$,
$\langle j, i, \mathsf{Concede}, \neg p_4 \wedge p_5 \to p_3\rangle$,
$\langle j, i, \mathsf{Argue}, (p_3 SO p_4)\rangle$,
$\langle j, i, \mathsf{Challenge}, p_3 \to p_1\rangle$

(vi) $\langle i, j, \mathsf{Resolve}, p_4\rangle$,
$\langle i, j, \mathsf{Argue}, (\neg p_4 SO p_3 \to p_1)\rangle$,
$\langle i, j, \mathsf{Challenge}, p_3 \to p_4\rangle$

(vii) $\langle j, i, \mathsf{WRetract}, p_4\rangle$,
$\langle j, i, \mathsf{WRetract}, p_3 \to p_4\rangle$,
$\langle j, i, \mathsf{SRetract}, p_5\rangle$,
$\langle j, i, \mathsf{SRetract}, p_3\rangle$,
$\langle j, i, \mathsf{WRetract}, p_4 \wedge p_5 \to p_2\rangle$,
$\langle j, i, \mathsf{Concede}, \neg p_4 \to (p_3 \to p_1)\rangle$,
$\langle j, i, \mathsf{Concede}, p_3 \to p_1\rangle$,
$\langle j, i, \mathsf{Concede}, p_1\rangle$,
$\langle j, i, \mathsf{Argue}, (p_6 SO p_2)\rangle$,

Fig. 3. Example of dialogue (see [19, p. 153])

are jointly inconsistent, in order to keep in line with standard properties of the modal operator G.

Now we can establish formally that our logic captures W&K's PPD_0-dialogues. For example we have:

Theorem 5.1

$$LAWS \models After_{\langle s,h,\mathsf{Assert},p\rangle}((\neg WC_h\, p \wedge \neg Done_{\langle h,s,\mathsf{Challenge},p\rangle}\top) \to G(Happens(\langle h, s, \mathsf{Challenge}, p\rangle) \vee Happens(\langle h, s, \mathsf{Concede}, p\rangle)))$$

Thus after an assertion of p the only possible reactions of the hearer are to either challenge or concede p, under the condition that he has not doubted that $\neg p$, and that he has not challenged p in the preceding move.

Proof. *LAWS* contains (see Table 2) the formula

$$After_{\langle s,h,\mathsf{Assert},p\rangle} G(Ch_h Happens(\langle h, s, \mathsf{Challenge}, p\rangle) \vee Ch_h Happens(\langle h, s, \mathsf{Concede}, p\rangle))$$

The precondition for $\langle h, s, \mathsf{Challenge}, p\rangle$ is

$$\neg WC_h\, p \wedge SC_s\, p \wedge \neg Done_{\langle h,s,\mathsf{Challenge},p\rangle}\top$$

Now the postcondition of $\langle s, h, \mathsf{Assert}, p\rangle$ is $SC_s\, p$. Hence we have by the law

Grounded propositions	SC_i	WC_i	SC_j	WC_j
$\emptyset$	p_1		p_2	
$WC_i\, p_3$ $SC_i\, p_3, SC_i\, p_3 \rightarrow p_1$	p_1, $p_3, p_3 \rightarrow p_1$		p_2	
$WC_j\, p_3 \rightarrow p_1, SC_j\, p_4,$ $SC_j\, p_5, SC_j\, p_4 \wedge p_5 \rightarrow p_2$			p_2, p_4, p_5 $p_4 \wedge p_5 \rightarrow p_2$	$p_3 \rightarrow p_1$
$WC_i\, p_5, WC_i\, p_4 \wedge p_5 \rightarrow p_2$ $SC_i \neg p_4, SC_i\, p_5,$ $SC_i \neg p_4 \wedge p_5 \rightarrow p_3$	$p_1, p_3, p_3 \rightarrow p_1$ $\neg p_4, p_5,$ $\neg p_4 \wedge p_5 \rightarrow p_3$	p_5 $p_4 \wedge p_5 \rightarrow p_2$		
$\neg SC_j\, p_3 \rightarrow p_1, WC_j\, p_3,$ $WC_j \neg p_4 \wedge p_5 \rightarrow p_3,$ $SC_j\, p_3, SC_j\, p_3 \rightarrow p_4,$ $WC_j \neg p_4$			p_2, p_4, p_5, p_3 $p_4 \wedge p_5 \rightarrow p_2,$ $p_3 \rightarrow p_4$	$\neg p_4$ $\neg p_4 \wedge p_5 \rightarrow p_3$
$SC_i \neg p_4,$ $SC_i \neg p_4 (\rightarrow p_3 \rightarrow p_1)$	$p_1, p_3, p_3 \rightarrow p_1$ $\neg p_4, p_5,$ $\neg p_4 \wedge p_5 \rightarrow p_3$ $\neg p_4 (\rightarrow p_3 \rightarrow p_1)$	p_5 $p_4 \wedge p_5 \rightarrow p_2$		
$\neg SC_j\, p_4, \neg WC_j\, p_4$ $\neg WC_j\, p_3 \rightarrow p_4, \neg SC_j\, p_3$ $\neg SC_j\, p_3 \rightarrow p_4, \neg SC_j\, p_5$ $\neg WC_j\, p_4 \wedge p_5 \rightarrow p_2,$ $\neg SC_j\, p_4 \wedge p_5 \rightarrow p_2$ $WC_j\, p_3 \rightarrow p_1, WC_j\, p_1$ $WC_j \neg p_4 \rightarrow (p_3 \rightarrow p_1)$ $SC_j\, p_6, SC_j\, p_6 \rightarrow p_2$			p_2 $p_6, p_6 \rightarrow p_2$	$\neg p_4$ $\neg p_4 \wedge p_5 \rightarrow p_3$ $p_3, p_5,$ $\neg p_4 \rightarrow (p_3 \rightarrow p_1)$ $p_3 \rightarrow p_1, p_1$

Table 3
Commitment stores in the example dialogue

of intentional action ($\text{Int}_{Ch_i,\alpha_i}$):

$$LAWS \models After_{\langle s,h,\mathsf{Assert},p\rangle}(\neg WC_h\, p \wedge \neg Done_{\langle h,s,\mathsf{Challenge},p\rangle} \top \rightarrow (Ch_h Happens(\langle h, s, \mathsf{Challenge}, p\rangle) \rightarrow Happens(\langle h, s, \mathsf{Challenge}, p\rangle)))$$

Similarly, for concede we have:

$$LAWS \models After_{\langle s,h,\mathsf{Assert},p\rangle}(\neg WC_h\, p \rightarrow (Ch_h Happens(\langle h, s, \mathsf{Concede}, p\rangle) \rightarrow Happens(\langle h, s, \mathsf{Concede}, p\rangle)))$$

Combining these two with the law of intentional action for Assert we obtain our theorem. □

Similar results for the other speech acts can be stated. They formally express and thus make more precise further properties of W&K's dialogue games. For example, the above theorem illustrates something that remained implicit in W&K's PPD_0 dialogues: the hearer of an assertion that p should not be committed that p himself because, if he were not, the dialogue would no more be a persuasion dialogue and no rule would apply.

Similarly, in a context where h's commitment store contains $SC_h\,(p \vee q)$, $SC_h\,\neg p$, and $SC_h\,\neg q$ (and is thus clearly inconsistent), W&K's dialogue rules do not allow s to execute $\langle s, h, \mathsf{Resolve}, p \vee q, \neg p \wedge \neg q\rangle$. This seems nevertheless a natural move in this context. Our formalization allows for it, the formal reason being that our logic of G is a normal modal logic, and thus validates $(SC_i\, p \wedge SC_i\, q) \rightarrow SC_i\,(p \wedge q)$.

6 Conclusion

The main contribution of this paper is the definition of a logic of grounding. We have shown that this notion has its origins in speech act theory [16,17], philosophy of mental states [11], and in philosophy of social action [15]. It is thus a philosophically well-founded notion.

Our formalisation is new as far as we are aware. Just as the structural approaches to dialogue it requires no hypotheses on the internal principles of the agents and accounts for the observation of a dialogue by a third party. Our characterization of speech acts is limited to the establishment of what must be true in order to avoid self-contradictions of the speaker.

We think that our work is very close of the notion of ostensible mental states of [7] and that our works could converge to very interesting results, for example on the definition of the semantics of a speech acts library.

Another feature of our notion is that it bridges the gap between mentalist and structural approaches to dialogue, by accounting for an objective viewpoint on dialogue by means of a logic involving belief.

We did not present a formal account of the dynamics. This requires the integration of a solution to the classical problems in reasoning about actions

(frame problem, ramification problem, and belief revision). These technical aspects will be described in future work.

Once we have such a formalism at our disposal it can be used to analyse dialogue corpora in order to formally derive whether some proposition is grounded or not for the participants. This could provide then an explanation for some cases of misunderstanding.

References

[1] Chellas, B. F., "Modal Logic: an introduction," Cambridge University Press, 1980.

[2] Cohen, P. R. and H. J. Levesque, *Intention is choice with commitment*, Artificial Intelligence Journal **42** (1990).

[3] Conte, R. and C. Castelfranchi, "Cognitive and Social Action," UCL Press, London, 1995.

[4] Fitting, M. C., "Proof Methods for Modal and Intuitionistic Logics," D. Reichel Publishing Company, Dordrecht, Netherlands, 1983.

[5] Fornara, N. and M. Colombetti, *Operational Specification of a Commitment-Based Agent Communication Language*, in: C. Castelfranchi and L. W. Johnson, editors, *Proc. First Int. Joint Conf. on Autonomous Agents and MultiAgent Systems (AAMAS-2002)*, ACM Press **2**, 2002, pp. 535–542.

[6] Herzig, A. and D. Longin, *C&L intention revisited*, in: D. Dubois, C. Welty and M.-A. Williams, editors, *Proc. 9th Int. Conf. on Principles on Principles of Knowledge Representation and Reasoning (KR2004) , Whistler, Canada* (2004), pp. 527–535.

[7] Nickles, M., F. Fischer and G. Weiss, *A framework for the representation of opinions and ostensible intentions*, To appear in this Volume.

[8] Sadek, M. D., *A study in the logic of intention*, in: B. Nebel, C. Rich and W. Swartout, editors, *Proc. Third Int. Conf. on Principles of Knowledge Representation and Reasoning (KR'92)* (1992), pp. 462–473.

[9] Searle, J. R., "Speech acts: An essay in the philosophy of language," Cambridge University Press, New York, 1969.

[10] Searle, J. R., "Expression and Meaning. Studies on the Theory of Speech Acts," Cambridge University Press, 1979.

[11] Searle, J. R., "Intentionality: An essay in the philosophy of mind," Cambridge University Press, 1983.

[12] Searle, J. R. and D. Vanderveken, "Foundation of illocutionary logic," Cambridge University Press, 1985.

[13] Singh, M. P., *Agent Communication Languages: Rethinking the Principles*, Computer **31** (1998), pp. 40–47.

[14] Singh, M. P., *A Social Semantics for Agent Communication Languages*, in: F. Dignum and M. Greaves, editors, *Issues in Agent Communication*, number 1916 in LNAI (2000), pp. 31–45.

[15] Tuomela, R., *Group beliefs*, Synthese **91** (1992), pp. 285–318.

[16] Vanderveken, D., "Principles of language use," Meaning and Speech Acts **1**, Cambridge University Press, 1990.

[17] Vanderveken, D., "Formal semantics of success and satisfaction," Meaning and Speech Acts **2**, Cambridge University Press, 1991.

[18] Verdicchio, M. and M. Colombetti, *A Logical Model of Social Commitment for Agent Communication*, in: *Proc. Second Int. Joint Conf. on Autonomous Agents and MultiAgent Systems (AAMAS-2003)* (2003), pp. 528–535.

[19] Walton, D. N. and E. C. Krabbe, "Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning," State University of New-York Press, NY, 1995.

A New Semantics for the FIPA Agent Communication Language based on Social Attitudes

Benoit Gaudou [1] and **Andreas Herzig** [1] and **Dominique Longin** [1] and **Matthias Nickles** [2]

Abstract. One of the most important aspects of the research on agent interaction is the definition of *agent communication languages* (ACLs), and the specification of a proper formal semantics of such languages is a crucial prerequisite for the usefulness and acceptance of artificial agency. Nevertheless, those ACLs which are still mostly used, especially the standard FIPA-ACL, have a communication act semantics in terms of the participating agents' *mental attitudes* (viz. beliefs and intentions), which are in general undeterminable from an external point of view due to agent autonomy. In contrast, semantics of ACLs based on *commitments* are fully verifiable, but not sufficiently formalized and understood yet. In order to overcome this situation, we propose a FIPA-ACL semantics which is fully verifiable, fully formalized, lean and easily applicable. It is based on *social attitudes* represented using a *logic of grounding* in straightforward extension of the BDI agent model.

1 Introduction

The design of agent communication languages (ACLs) has attracted a lot of attention during the last years. Such languages are mostly based on Searle and Vanderveken's speech act theory [9], and are not only relevant for applications involving real software agents or robots, but also for other software entities which need to communicate, like web services.

Among the different existing ACLs, FIPA-ACL is still the most important standard, subsets of which are widely used in agent interaction protocols. FIPA-ACL is semantically rich, and the concepts involved are quite intuitive.

Nevertheless, FIPA-ACL has a feature that has often been criticized in the literature, viz. that the semantics of communication acts (CAs) is defined in terms of the agents' mental states. For example, when agent i informs agent j that φ, then the (rational) effect is that agent j starts to believe φ. In order for such an effect to obtain some hypotheses have to be made; but even in such contexts j is autonomous and might not adopt φ, and in any case i or other agents and the system designer can never verify whether this is the case or not. This is especially felt as being too strong in open environments with black- or gray-box agents where we don't even want to ascribe mental attitudes to other agents.

In contrast, those semantics based on the concept of social commitments [10] is verifiable because they are only based on what has been communicated and the commitments the agents have made by doing that (instead of the beliefs and intentions that are "behind" these commitments and that have caused them). The drawback here is that the existing approaches are only semi-formal, in particular because there is no consensus on what "being committed" actually means. As a consequence, they are rarely used in practice up to now.

The aim of this paper is to resolve the problems of FIPA's CA semantics without loosing its benefits. We propose a novel semantics avoiding the strong hypotheses of the original semantics by "lifting" the BDI-based FIPA semantics to the social level. We do so by replacing the usual private mental attitudes of BDI logics by *public* mental attitudes, i.e. attitudes that have been made public through communication (*social attitudes*). More precisely, our semantics is based on an unified and extended approach to the concept of *communication attitudes (ostensible beliefs and intentions)* [7] and the more or less equivalent concept of *grounding*[3] [5]. For example, the effect of an informing-that-p act is that it is public that the sender believes that p. This does not mean that the sender really believes that p, but only hinders him to subsequently inform that $\neg p$, or to inform that he ignores whether p.

The major benefits of our new semantics are the following:

- It is verifiable, and suitable even for truly autonomous, possibly malevolent agents.
- It is fully formalized.
- It is based on a straightforward extension of BDI, and therefore relatively lightweight.
- It can easily be adapted to similar ACLs, e.g. the widely used KQML/KIF.
- It generalizes the single-addressee FIPA acts to groups of agents, and it distinguishes the group of addressees from the group of bystanders (overhearers), and thus refines FIPA's acts.

All in all, we aim at an agent communication semantics that eliminates the major shortcomings of the still predominant mentalist approaches to ACL semantics while being "upward compatible" to the standard FIPA semantics and similar approaches, lean, and formally well founded.

The remainder of this paper is organized as follows: The next section provides a short account of the logical framework that we have chosen as a formal foundation of our approach. Section 3 presents the new semantics, and Section 4 illustrates our approach by means of a case study. Section 5 concludes.

2 A logic of grounding

In this section we briefly present the logic of belief, intention, action, and grounding defined in [5], that is based on Cohen and

[1] Université Paul Sabatier, IRIT, Toulouse, email: gaudou,herzig,longin@irit.fr

[2] Technical University of Munich, email: nickles@informatik.tu-muenchen.de

[3] We use the term *grounding* as Traum [11], *i.e.* it refers to "the process of adding to the common ground between conversational agent".

Levesque's [2]. AGT is a finite set of agents, ACT is a set of actions, $ATM = \{p, q, ...\}$ is the set of atomic formulas. Complex formulas are denoted by φ, ψ... A model $\mathcal{M}$ is a 5-tuple that is made up of: a set of possible worlds W; a mapping

$$\mathcal{V} : W \longrightarrow (ATM \longrightarrow \{0,1\})$$

associating a valuation $\mathcal{V}_w$ to every $w \in W$; a mapping

$$\mathcal{A} : ACT \longrightarrow (W \longrightarrow 2^W)$$

associating actions $\alpha \in ACT$ and worlds $w \in W$ with the set of worlds resulting from the execution of α in w; a mapping

$$\mathcal{G} : (2^{AGT} \setminus \emptyset) \longrightarrow (W \longrightarrow 2^W)$$

associating sets of agents $I \subseteq AGT$ and worlds $w \in W$ with the set of worlds that are publicly possible for the group I at w (the worlds that are compatible with what has been uttered in I's presence); and finally the mapping

$$\mathcal{I} : AGT \longrightarrow (W \longrightarrow 2^{2^W})$$

associating every $i \in AGT$ and world w with the set of propositions (alias sets of worlds) that are intended by i. (The $\mathcal{I}_i$ are neighborhood functions in Chellas' sense [1].)

The logical language contains modal operators of action $After_\alpha$ and $Before_\alpha$, for every $\alpha \in ACT$, modal operators of groundedness G_I for every group I, and modal operators of intention Int_i for every agent $i \in AGT$.

The formula $After_\alpha\, \varphi$ reads "φ is true after every execution of the action α", and $Before_\alpha\, \varphi$ reads "φ is true before every execution of the action α". Semantically, $w \Vdash After_\alpha\, \varphi$ iff $w' \Vdash \varphi$ for each $w' \in \mathcal{A}_\alpha(w)$, and $w \Vdash Before_\alpha\, \varphi$ iff $w' \Vdash \varphi$ for each w' such that $w \in \mathcal{A}_\alpha(w')$. The logic of $After_\alpha$ and $Before_\alpha$ is the tense logic K^t, i.e. standard normal modal logic K plus the conversion axioms $\varphi \rightarrow Before_\alpha \neg After_\alpha \neg\varphi$ and $\varphi \rightarrow After_\alpha \neg Before_\alpha \neg\varphi$. The abbreviation $Done_\alpha\, \varphi \stackrel{def}{=} \neg Before_\alpha \neg\varphi$ reads "α has just been done before which φ was true". We note $Done(\alpha) \stackrel{def}{=} Done_\alpha \top$ for convenience. Moreover, $Before_{\alpha\cup\alpha'}\, \varphi$ abbreviates $Before_\alpha\, \varphi \wedge Before_{\alpha'}\, \varphi$. (Hence $Done(\alpha\cup\alpha')$ stands for $Done(\alpha)\vee Done(\alpha')$.)

$G_I\, \varphi$ reads "it is grounded for group I that φ is true", or for short: "φ is grounded for I". When I is a singleton, $G_{\{i\}}\, \varphi$ means that for agent i, φ is grounded. In this (and only in this) degenerated case 'public' grounding is the same as private belief. We write $G_i\, \varphi$ for $G_{\{i\}}\, \varphi$. The accessibility relations of grounding operators must satisfy the constraints for the standard normal modal logic KD (seriality), plus the following, for groups I, I' such that $I' \subseteq I$:

(i) if $u\mathcal{G}_{I'}v$ and $v\mathcal{G}_I w$ then $u\mathcal{G}_I w$
(ii) if $u\mathcal{G}_{I'}v$ and $u\mathcal{G}_I w$ then $v\,G_I\, w$
(iii) if $u\mathcal{G}_I v$ and $v\mathcal{G}_{I'}w_1$ then there is w_2 such that $u\mathcal{G}_I w_2$ and $V(w_1) = V(w_2)$
(iv) $\mathcal{G}_I \subseteq \bigcup_{i\in I} \mathcal{G}_I \circ \mathcal{G}_i$

Constraint (i) stipulates that subgroups are aware of what is grounded in the group: whenever w is a world for which it is grounded for I' that all I-grounded propositions hold in w, then all I-grounded propositions indeed hold in w. This is a kind of *attention* property: each subgroup participating in a conversation is aware of what is grounded in the group. Similarly (ii) expresses that subgroups are aware of what is ungrounded in the group, too. (i) and (ii) correspond to the axioms of strong rationality (SR_+) and (SR_-):

$$G_I\, \varphi \rightarrow G_{I'}\, G_I\, \varphi \qquad (SR_+)$$

$$\neg G_I\, \varphi \rightarrow G_{I'}\, \neg G_I\, \varphi \qquad (SR_-)$$

which express that if a proposition φ is grounded (resp. ungrounded) for a group I then it is grounded for each subgroup that φ is grounded (resp. ungrounded) for I[4].

(iii) stipulates that for every objective proposition grounded for I it is publicly established for I that each subgroup of I is grounded on it (which does not imply that it is grounded for the latter): whenever w is a world for which all propositions grounded for I' are grounded for I, then all those propositions are indeed grounded for I in w. It validates the axiom (WR)

$$G_I\, \varphi \rightarrow G_I\, G_{I'}\, \varphi, \text{ for } \varphi \text{ objective} \qquad (WR)$$

which says that if the objective formula φ is grounded for a group K then it is necessarily grounded for K that for each subgroup K' the formula is grounded.[5] Note that this does not imply that for every subgroup φ is actually grounded, *i.e.* (WR) does not entail $G_K\, \varphi \rightarrow G_{K'}\, \varphi$. In particular, the fact that φ is grounded for group K does not imply that the members of K believe that φ.

(iv) expresses that if it is grounded for a set I that a proposition is established for every agent then it is grounded for I, too. This corresponds to axiom (CG)

$$(\bigwedge_{i\in I} G_I\, G_i\, \varphi) \rightarrow G_I\, \varphi \qquad (CG)$$

which says that if a proposition is established for every agent in I, then it is established for the whole group I. Together, (WR) and (CG) stipulate that for objective φ we have $(\bigwedge_{k\in K} G_K\, G_k\, \varphi) \leftrightarrow G_K\, \varphi$. Note that $G_K\, \varphi$ does NOT imply $G_k\, \varphi$ where $k \in K$. Indeed, a proposition can be grounded in a group independently of the private belief of each agent of the group about this proposition: there is thus no sincerity hypothesis.

$Int_i\, \varphi$ reads "agent i intends that φ be true". The Int_i are non-normal modal operators which only validate the rule of equivalence: $\frac{\varphi\leftrightarrow\psi}{Int_i\, \varphi\leftrightarrow Int_i\, \psi}$. They neither validate $Int_i\, (\varphi \wedge \varphi') \rightarrow (Int_i\, \varphi \wedge Int_i\, \varphi')$ nor $(Int_i\, \varphi \wedge Int_i\, \varphi') \rightarrow Int_i\, (\varphi \wedge \varphi')$.

Intentions and actions are related by the principle of intentional action saying that if α has just been performed by agent i then i had the intention to do so immediately before.

$$Before_{i:\alpha}\, Int_i\, Done(i{:}\alpha) \qquad (IA)$$

where $i{:}\alpha$ denotes that action α is performed by agent i.

To highlight our proposal for the semantics of grounding consider the following example. There are three agents $AGT = \{0, 1, 2\}$. Let agent 0 (privately) believe that 2 sells high-quality products, formally written $G_0\, q_2$. Now suppose that in private conversation agent 0 tells 1 that the contrary is the case (for example to trigger some attitude of 1 that benefits 0). The (illocutionary) effect is $G_{\{0,1\}}\, G_0\, \neg q_2$. Then agent 2 joins in the conversation, and later on 0 informs 1 and 2 that q_2: The illocutionary effect is $G_{\{0,1,2\}}\, G_0\, q_2$. This illustrates that even for nested groups $\{0\} \subset \{0, 1\} \subset \{0, 1, 2\}$, mutually inconsistent states of public group belief might hold simultaneously.

3 Communication act semantics

Following and extending the ACL syntax used in [4], a single communication act (CA) is denoted as $\langle i, \mathsf{ActName}(J, \varphi), K\rangle$, where i

[4] In particular, we have the modal axioms (4) and (5) for G_I operators as theorems of our logic.

[5] (WR) concerns only objective formulas, i.e. formula that does not contain any modality. If we applied (WR) to some mental states, we would restrict the agents' autonomy. For example, when an agent performs the speech act $\langle i, \mathsf{Inform}(J, p), K\rangle$, he expresses publicly that he believes p. Thus if agent i expresses: $\langle i, \mathsf{Inform}(J, G_J\, p), K\rangle$ the formula $G_K\, G_i\, G_J\, p$ holds, and the agents $j \in J$ cannot afterwards express that they believe $\neg p$. If he made this speech acts, the formulae $G_K\, G_J\, \neg p$ and, thanks to (WR), $G_K\, G_i\, G_J\, \neg p$ would hold, which is inconsistent with the above formula $G_K\, G_i\, G_J\, p$.

is the performing agent, J is a group of recipients (whereas FIPA only allows one addressee). *ActName* is the name of the act (in our model not necessarily corresponding to exactly one speech act type, see below). φ is the propositional content of the act. K, which is missing in FIPA, denotes a group of attending agents who overhear the respective utterance, with $i \in K$, $J \subseteq K \setminus \{i\}$ and $J \neq \emptyset$. For a dialogue of only two agents i and j we have $J = \{j\}$ and $K = \{i, j\}$.

In the standard semantics of FIPA CAs [4] (henceforth called FIPA-S), semantics is specified by providing the *feasibility preconditions* (FPs) and the *rational effects* (REs) of single CAs. The former denote which logical conditions need to be fulfilled in order to execute the respective act, and the latter specify which conditions hold after the successful performance of that act. FPs characterize both the ability of the speaker to perform the act and the context-dependent relevance of the act (i.e., that performing the act is relevant given a certain dialogue context). In contrast, REs specify the desired and rationally-expectable direct perlocutionary effect of the utterance, i.e. what becomes true in case the perlocutionary act succeeds.

We think there are at least three reasons not to qualify a CA by its rational effect. Firstly, it is possible to desire and expect different kinds of RE of the same CA; secondly, Searle shows in [9, Sec. 2.1] that the effect of a speech act cannot be a rational (or perlocutionary) effect simply because a lot of speech acts just do not have any perlocutionary effect. He also shows that even if a speech act can have a perlocutionary effect, we can always exhibit a context where the speaker does not intend this perlocutionary effect. Thirdly, strong hypotheses (such as sincerity, competence, credibility…) must be made about the agents to enable the inference of the expected RE, which is too restrictive in our context of open multi-agent systems, possibly with conflicts and malevolent, egocentric agents...

In contrast to FIPA-S, the FPs and IEs (for illocutionary effects) in our model do not make any statement about mental attitudes, but specify the preconditions and effects in terms of groundings of group K (the public, so to say). They are chosen such that the respective communication act is both executable given all realistic preconditions, and succeeds reliably with a publicly verifiable effect. The only (self-evident) exception follows from the bridge axioms (SR_+) and (SR_-) given in the previous section, stating that an agent or subgroup of a certain group knows about the existence of the respective grounded beliefs or intentions of their group — this means merely that the agents keep track of the ongoing course of communication in terms of FPs and IEs.

In the sequel we use the term *Social Attitudes Based Semantics* (SABS) for our modelling, and will define the SABS semantics of the four *primitive CAs* of FIPA-ACL: *Inform*, *Request*, *Confirm* and *Disconfirm*, and we will also present the respective FIPA-S specifications for comparison. All other FIPA-CAs are macros composed of these primitives in a more or less straightforward manner.

3.1 Inform: Asserting information

We start with the FIPA-S version of the semantics:

$\langle i, \mathsf{Inform}_{\mathsf{FIPA}}(j, \varphi), K\rangle$

FP: $Bel_i\,\varphi \wedge \neg Bel_i\,(Bel_j\,\varphi \vee Bel_j\,\neg\varphi \vee U_j\varphi \vee U_j\neg\varphi)$

RE: $Bel_j\,\varphi$

At this, $U_j\varphi$ denotes that agent j is uncertain about φ, but thinks that φ is more likely than $\neg\varphi$. The terms "uncertain" and "more likely" are not precisely defined in [4]. The essential preconditions of *Inform* in the FIPA-S are thus that agent i truthfully believes what he asserts, and that the receiver does not have any definite opinion about the asserted proposition. Whereas the former condition is obviously unrealistic given a truly autonomous agent i, the latter disallows (problematically) the use of *Inform* to convince the addressee. We consider the latter usage as crucial e.g. in the context of computational argumentation and argumentation-based negotiation. We could introduce an additional conviction-act extending the syntax, or try to emulate it with a construct like $Request(Inform(\varphi))$, but this would not only be unnecessary and inelegant, but would also blur the fact that there exists a continual transition from "pure information" to conviction. It is also not clear why the absence of an opinion shall be a realistic precondition for the success of an information act, or, conversely, why the existence of an opinion (which could be very weak, or "by default" only) shall hinder the receiver to adopt the asserted information (e.g., consider that the addressee might trust the sender more than herself).

The rational effect of $Inform$ in FIPA-S is simply that the addressed agent believes what she has been told (in case the act succeeds). Of course, this effect cannot be verified with autonomous agents. Even if it could be verified, it would be too strong and unlikely. Moreover it is not verifiable that the informing agent (truthfully) intends the adoption of a certain belief.

These concerns lead to the following SABS semantics:

$\langle i, \mathsf{Inform}(J, \varphi), K\rangle$

FP: $\neg G_K\; G_J\,\varphi \wedge \neg G_K\; Int_i\; G_J\,\varphi \wedge \neg G_K\,\neg G_i\,\varphi$

IE: $G_K\; G_i\,\varphi \wedge G_K\; Int_i\; G_J\,\varphi$

In the FP, $\neg G_K\; G_J\,\varphi$ specifies that the addressed agent has not expressed φ before (with group K attending), corresponding to the $\neg Bel_i\; Bel_j\,\varphi$ part of the FIPA-S FP (the *relevance precondition*). It simply expresses that asserting an information would be unnecessary if the receiver has already expressed its belief in it. However, our new FP does not demand that group J has no opinion at all about φ, allowing to use $Inform$ also to *convince* J in case this group has already expressed its disbelief in φ earlier. $\neg G_K\; Int_i\; G_J\,\varphi$ in FP effectively demands that agent i did not assert this information using an assertive communication act before, which is also an aspect of the relevance precondition. $\neg G_K\,\neg G_i\,\varphi$ ensures that the asserted opinions of agent i are mutually consistent. (The latter is a precondition of *rationality*).

In the IE, $G_K\; G_i\,\varphi$ denotes that with asserting φ, it becomes grounded that agent i believes that φ, regardless if she does so privately (i.e., mentally) or not.

As usual we define $\langle i, \mathsf{InformIf}(J, p), K\rangle$ as an abbreviation of $\langle i, \mathsf{Inform}(J, p), K\rangle \cup \langle i, \mathsf{Inform}(J, \neg p), K\rangle$. Hence $Done(\langle i, \mathsf{InformIf}(J, p), K\rangle) \equiv Done(\langle i, \mathsf{Inform}(J, p), K\rangle) \vee Done(\langle i, \mathsf{Inform}(J, \neg p), K\rangle)$.

However, in many cases, we can safely assume that group J immediately starts to publicly believe the asserted information, namely when this group apparently trusts the uttering agent in regard to this information. (A notorious exception are exam situations.) An important particular case is expressed by the following axiom, for $J \subseteq K$ and $\alpha = \langle i, \mathsf{InformIf}(J, \varphi), K\rangle$:[6]

$$(G_K\; Done_\alpha \bigwedge_{j \in J} Int_j\; Done(\alpha)) \rightarrow G_K\; G_J\,\varphi \qquad (1)$$

This specifies that if an agent has requested a certain information before from agent i in form of a closed question (like with "Is it raining

[6] We here consider that the group J have asked i to publicly declare that φ. (And not each j, as it would be the case if α was $\langle i, \mathsf{InformIf}(\{j\}, \varphi), K\rangle$ in (1).)

outside?"), it becomes grounded that she believes the answer.[7]

3.2 Request: Requesting an action to be done

Again, we state the FIPA version of the semantics first:

$\langle i, \mathsf{Request}_{\mathsf{FIPA}}(j, \alpha), K\rangle$

FP: $FP(\alpha)[i\backslash j] \wedge Bel_i\, Agent(j, \alpha) \wedge Bel_i \neg PG_j\, Done(\alpha)$

RE: $Done(\alpha)$

Here, α is an action expression, $FP(\alpha)[i\backslash j]$ denotes the part of the feasibility preconditions of action α where the mental attitudes are those of agent i. $Agent(j, \alpha)$ states that j is the only agent that ever performs, has performed or will perform α, and $PG_j\, Done(\alpha)$ denotes that $Done(\alpha)$ (i.e., action α has just been performed successfully) is a *persistent goal* [8] of agent j. The RE just specifies that the intended perlocutionary effect of this communication act is to get α done.

Obviously, these specifications again require strong assumptions about mental properties, which are equally problematic as in the case of $Inform$. In addition, $Agent(j, \alpha)$ reduces the scope of this communication act unnecessarily, disallowing concurrent intention of j to perform the same action herself.

As in our formalism the propositional content of a CA is a formula, a request to do action α is defined as a request that $Done(\alpha)$ be true. Furthermore, in our case the addressee of a speech act is a group of agents. Thus a request is addressed to each agent of the group in the aim that either at least one agent of the group do the requested action ("open that door"), or each agent of the group do it ("clean that room", addressed to a group of children). $i : \alpha$: denotes that i is the author of action α (making superfluous the FIPA *Agent* predicate). We thus have two kinds of request (whereas there is only one in FIPA):

$\langle i, \mathsf{RequestSome}(J, J{:}\alpha), K\rangle \stackrel{def}{=}$

$$\langle i, \mathsf{RequestSome}(J, \bigvee_{j\in J} Done(j{:}\alpha)), K\rangle$$

FP: $\left(\neg G_K \bigvee_{j\in J} Int_j\, Done(j{:}\alpha)\right) \wedge \neg G_K \neg Int_i \left(\bigvee_{j\in J} Done(j{:}\alpha)\right)$

IE: $G_K\, Int_i \left(\bigvee_{j\in J} Done(j{:}\alpha)\right) \wedge G_K \neg G_i \left(\bigvee_{j\in J} Int_j\, Done(j{:}\alpha)\right)$

So our FP specifies that is not grounded that at least one of the agents in J intends to achieve α already (relevance precondition), and that it is not grounded that agent i does not intend $Done(\alpha)$ (rationality precondition). The IE is also straightforward: the act results in the grounding that agent i intends that at least one agent in J intends $Done(\alpha)$ become true, and that i does not believe that one agent in J intends $Done(\alpha)$.

Second, we define:

$\langle i, \mathsf{RequestEach}(J, J{:}\alpha), K\rangle \stackrel{def}{=}$

$$\langle i, \mathsf{RequestEach}(J, \bigwedge_{j\in J} Done(j{:}\alpha)), K\rangle$$

FP: $\left(\neg G_K \bigwedge_{j\in J} Int_j\, Done(j{:}\alpha)\right) \wedge \neg G_K \neg Int_i \left(\bigwedge_{j\in J} Done(j{:}\alpha)\right)$

IE: $G_K\, Int_i \left(\bigwedge_{j\in J} Done(j{:}\alpha)\right) \wedge G_K \neg G_i \left(\bigwedge_{j\in J} Int_j\, Done(j{:}\alpha)\right)$

which specifies that i intends that each agent of J perform the requested action α. For compatibility reasons, we also define

$\langle i, \mathsf{Request}(J, \alpha), K\rangle \stackrel{def}{=}$

$\langle i, \mathsf{RequestSome}(J, J{:}\alpha), K\rangle$

FIPA also defines the acts $Confirm$ (for the confirmation of an uncertain information) and its pendant $Disconfirm$ as primitives. But since our $Inform$ semantics has an adequately weakened FP that does not require that the asserted information is not uncertain, $Confirm$ and $Disconfirm$ simply map to $Inform$ in our semantics.

4 Case study

In order to demonstrate the properties and the application of our approach, this section presents a brief case study in form of an *agent purchase negotiation* scenario. In particular, we aim to demonstrate the following crucial features of SABS, all not being present in FIPA-S or, by principle, any other BDI-based ACL semantics:

- Pre- and post-conditions of communication acts being only dependent from publicly observable agent behavior, thus being fully verifiable;
- Communication acts with contents being inconsistent with the beliefs and intentions of the participating agents;
- Communication acts addressing groups of agents;
- Multiple communication acts uttered by the same sender, but with mutually inconsistent contents (even towards nested groups);
- Persuasive Inform-acts.

In addition, the example shows how the logging of the grounding state of the negotiation dialogue can replace *commitment stores*, which are usually used to keep track of the various commitments arising during the course of an interaction (like to sell or buy a product). In contrast, by the use of our semantics we obtain the publicly available information about the state of commitment of the participating agents directly in terms of logical post-conditions of communication acts, namely publicly expressed intentions. As explained in Section 1, we consider this to be simpler and formally clear compared to the use of social commitments in the sense of [10].

The interaction roughly follows protocols for *purchase negotiation dialogue games* as known from, e.g., [6], but omitting several details of such protocols which are not relevant for our demonstrative purposes (like the specification of selling options in detail). Also, such protocols often make use of proprietary negotiation locutions, whereas we get along with FIPA-ACL constructs, since in our context, no acts not contained in FIPA-ACL (like the "Promise" and "Threaten" acts in protocols for argumentation-based negotiation) are required. Nevertheless, our scenario is clearly beyond FIPA's *contract net* specification [3].

Our scenario consists of four agents $MAS = \{s_1, s_2, b_1, b_2\}$, representing potential car sellers and customers. In the discourse universe exists two instances θ_1 and θ_2 of some car type θ (e.g., specimen of the Alfa Romeo 159).

We present now the interaction course, consisting of sequential steps in the following form. Note that the interaction course consists of multiple interlaced conversations among different sender/receiver pairs and different overhearers (i.e., different "publics" so to say). In particular, agent b_2 is involved in two selling dialogues at the same time.

Utterance no. sender→receiver: Descriptive act title
Message [8]

[7] The intention Int_j can be triggered with FIPA's *QueryIf* act. The schema would work analogously for $\langle i, \mathsf{InformIf}(\{j\}, \neg\varphi), K\rangle$.

[8] Using syntactical macros according to [4]. Only in case the message primitives are semantically relevant in our context, the respective macros are expanded.

Effect (optionally) gives the effect of the act in terms of grounded formulas, according to SABS and the axioms in Section 2 (so this may go beyond the direct IE).

In contrast, *Private information (PI)* optionally unveils relevant mental attitudes before or after an act has been uttered and understood by the respective agents. The PIs are not determined by preceding communication acts, due to agent autonomy. They are also of course usually not available to observers, and just given for explanatory purposes.

U1 $s_1 \rightarrow \{b_1, b_2\}$**: Initialize dialogue**
$\langle s_1, \mathsf{RequestEach}(\{b_1, b_2\}, enterDialogue(\theta_1)), \{s_1, b_1, b_2\}\rangle$
U2 $b_1 \rightarrow \{s_1\}$**: Enter dialogue**
$\langle b_1, \mathsf{Agree}(\{s_1\}, enterDialogue(\theta_1)), \{s_1, b_1, b_2\}\rangle$
U3 $b_2 \rightarrow \{s_1\}$**: Enter dialogue**
$\langle b_2, \mathsf{Agree}(\{s_1\}, enterDialogue(\theta_1)), \{s_1, b_1, b_2\}\rangle$
U4 $s_2 \rightarrow \{b_2\}$**: Initialize dialogue**
$\langle s_2, \mathsf{Request}(\{b_2\}, enterDialogue(\theta_2)), \{s_2, b_2\}\rangle$
U5 $b_2 \rightarrow \{s_2\}$**: Enter dialogue**
$\langle b_2, \mathsf{Agree}(\{s_2\}, enterDialogue(\theta_2)), \{s_2, b_2\}\rangle$
PI_{s_1}**:** $Bel_{s_1}\, discounts$
U6 $s_1 \rightarrow \{b_1, b_2\}$**: Information about discount**
$\langle s_1, \mathsf{Inform}(\{b_1, b_2\}, \neg discounts), \{s_1, b_1, b_2\}\rangle$
Effect:
$G_{\{s_1,b_1,b_2\}} G_{s_1} \neg discounts$
$\wedge G_{\{s_1,b_1,b_2\}} Int_{s_1} G_{\{b_1,b_2\}} \neg discount$
Seller s_1 asserts that no discounts can be given while believing (PI_{s_1}: $Bel_{s_1}\, discount$) that the opposite is true (there might be the company policy that discounts should be given, but that might reduce the seller's individual profit).
U7 $s_1 \rightarrow \{b_2\}$**: Information about discount**
$\langle s_1, \mathsf{Inform}(\{b_2\}, discounts), \{s_1, b_2\}\rangle$
Effect:
$G_{\{s_1,b_2\}} G_{s_1} discounts$
$\wedge G_{\{s_1,b_2\}} Int_{s_1} G_{b_2} discount$
While seller s_1 informed group $\{b_1, b_2\}$ that there would be no price discounts, he informs customer b_2 that this is not true (likely because s_1 thinks that b_2 is a valued customer whereas b_1 is not).
U8 $b_2 \rightarrow \{s_1\}$**: Query if car type has high accident rate**
$\langle b_2, \mathsf{Request}(\{s_1\}, InformIfAccidentRateHigh), \{s_1, b_2\}\rangle$
Effect:
$G_{\{s_1,b_2\}} Int_{b_2} Done(s_1 : InformIfAccidentRateHigh) \wedge$
$G_{\{s_1,b_2\}} \neg G_{b_2} Int_{s_1} Done(s_1 : InformIfAccidentRateHigh)$,
with
$InformIfAccidentRateHigh \stackrel{def}{=}$
$\langle s_1, \mathsf{InformIf}(\{b_2\}, accidentRateHigh(\theta)), \{s_1, b_2\}\rangle$
$PI_{s_1} : Bel_{s_1}\, accidentRateHigh(\theta_1)$
U9 $s_1 \rightarrow \{b_2\}$**: Information about accident rate**
$\langle s_1, \mathsf{Inform}(\{b_2\}, \neg accidentRateHigh(\theta)), \{s_1, b_2\}\rangle$
Effect:
$G_{\{s_1,b_2\}} G_{s_1} \neg accidentRateHigh(\theta)$
$\wedge G_{\{s_1,b_2\}} G_{b_2} \neg accidentRateHigh(\theta)$
Note that due to her closed question before and axiom 1 it becomes immediately grounded that b_2 believes the asserted information. In addition, b_2 privately believes this information also (see PI_{b_2} below), but revises this later.
Seller s_1 asserted $\neg accidentRateHigh(\theta_1)$ though thinking the opposite.
$PI_{b_2} : Bel_{b_2}\, \neg accidentRateHigh(\theta)$
U10 $b_2 \rightarrow \{s_2\}$**: Query if car type has high accident rate**
$\langle b_2, \mathsf{Request}(\{s_2\}, InformIfAccidentRateHigh), \{s_2, b_2\}\rangle$
U11 $s_2 \rightarrow \{b_2\}$**: Information about accident-damage**
$\langle s_2, \mathsf{Inform}(\{b_2\}, accidentRateHigh(\theta)), \{s_2, b_2\}\rangle$
Again, b_2 publicly believes the information, and trusts it for some reason privately more than the information given by seller s_1 earlier. Nevertheless, $G_{\{s_1,b_2\}} G_{b_2} \neg accidentRateHigh(\theta_1)$ remains true.
$PI_{b_2} : Bel_{b_2}\, accidentRateHigh(\theta)$
U12 $b_2 \rightarrow \{s_2\}$**: Propose to buy at a certain price**
$\langle b_2, \mathsf{Propose}(\{s_2\}, buy(\theta_2, 10000\pounds)), \{s_2, b_2\}\rangle$
U13 $s_2 \rightarrow \{b_2\}$**: Accept proposal**
$\langle s_2, \mathsf{AcceptProposal}(\{b_2\}, buy(\theta_2, 10000\pounds)), \{s_2, b_2\}\rangle$
Effect (with the previous act):
$G_{\{s_2,b_2\}} Int_{b_2} buy(\theta_2, 10000\pounds)$ (i.e., b_2 is publicly committed to buy θ_2 at the price of $10000\pounds$ now).

5 Conclusion

We've proposed a novel approach to the semantics of agent communication, based on verifiable social attitudes which are triggered by observable communication acts. We believe that this approach is more adequate for open systems in comparison both to traditional mentalistic and commitment-based semantics, as it allows to analyze the meaning of messages on the social level without the need to know about mental agent properties or architectural details, while being easily comprehensible, downward compatible to BDI, and fully formalized. A subject of future work in this respect will be the practical application of our approach in the field of interaction protocols, and argumentation and negotiation frameworks.

REFERENCES

[1] B. F. Chellas, *Modal Logic: an introduction*, Camb. Univ. Press, 1980.
[2] Philip R. Cohen and Hector J. Levesque, 'Intention is choice with commitment', *Artificial Intelligence*, **42**(2–3), 213–261, (1990).
[3] Foundation for Intelligent Physical Agents. FIPA Interaction Protocol Library Specification, 2000. URL: http://www.fipa.org/specs/fipa00025.
[4] Foundation for Intelligent Physical Agents. FIPA Communicative Act Library Specification, 2002. URL: http://www.fipa.org/repository/aclspecs.html.
[5] B. Gaudou, A. Herzig, and D. Longin, 'Grounding and the expression of belief', in *Proc. 10th Int. Conf. on Princ. of Knowledge Repr. and Reasoning (KR 2006)*, (2006).
[6] P. McBurney, R. M. van Eijk, S. Parsons, and L. Amgoud, 'A dialogue-game protocol for agent purchase negotiations', *Journal of Autonomous Agents and Multi-Agent Systems*, **7**(3), 235–273, (2003).
[7] M. Nickles, F. Fischer, and G. Weiss, 'Communication attitudes: A formal approach to ostensible intentions, and individual and group opinions', in *Proceedings of the Third International Workshop on Logic and Communication in Multiagent Systems (LCMAS 2005)*, eds., W. van der Hoek and M. Wooldridge, (2005).
[8] M. D. Sadek, 'A study in the logic of intention', in *Proc. Third Int. Conf. on Principles of Knowledge Representation and Reasoning (KR'92)*, eds., Bernhard Nebel, Charles Rich, and William Swartout, pp. 462–473. Morgan Kaufmann Publishers, (1992).
[9] J. R. Searle, *Speech acts: An essay in the philosophy of language*, Cambridge University Press, New York, 1969.
[10] M. P. Singh, 'A social semantics for agent communication languages', in *Proceedings of the IJCAI Workshop on Agent Communication Languages*, (2000).
[11] D. R. Traum, *Computational theory of grounding in natural language conversation*, Ph.D. dissertation, Computer Science Departement, University of Rochester, December 1994.

The logic of acceptance: grounding institutions on agents' attitudes

Emiliano Lorini, Dominique Longin, Benoit Gaudou, Andreas Herzig
Institut de Recherche en Informatique de Toulouse (IRIT)
118 Route de Narbonne, F-31062, Toulouse, France
{lorini,longin,gaudou,herzig}@irit.fr

January 23, 2009

Abstract

In the recent years, several formal approaches to the specification of normative multi-agent systems and artificial institutions have been proposed. The aim of this paper is to advance the state of the art in this area by proposing an approach in which a normative multi-agent system is conceived to be autonomous, in the sense that it is able to create, maintain, and eventually change its own institutions by itself, without the intervention of an external designer in this process. In our approach the existence and the dynamics of an institution (norms, rules, institutional facts, *etc.*) are determined by the (individual and collective) *acceptances* of its members, and its dynamics depends on the dynamics of these acceptances.

In order to meet this objective, we propose the logic $\mathcal{AL}$ (*Acceptance Logic*) in which the acceptance of a proposition by the agents *qua* members of an institution is introduced. Such propositions are true w.r.t. an institutional context and correspond to facts that are instituted in an attitude-dependent way.

The second part of the paper is devoted to the logical characterization of some important notions in the theory of institutions. We provide a formalization of the concept of *constitutive rule*, expressed by a statement of the form "X counts as Y in the context of institution x". Then, we formalize the concepts of obligation and permission (so called *regulative rules*). In our approach constitutive rules and regulative rules of a certain institution are attitude-dependent facts which are grounded on the acceptances of the members of the institution.

Keywords

Modal logic, institutions, acceptance, normative systems, multi-agent systems

1 Introduction

The problem of devising artificial institutions and modeling their dynamics is a fundamental problem in the multi-agent system (MAS) domain [Dignum and Dignum, 2001]. Following [North, 1990, p. 3], artificial institutions can be conceived as "the rules of the game in a society or the humanly devised constraints that structure agents' interaction". Starting from this concept of institution, many researchers working in the field of normative MAS have been interested in developing models which describe the different kinds of rules and norms that agents have to deal with. In some models of artificial institutions norms are conceived as means to achieve coordination among agents and agents are supposed to comply with them and to obey the authorities of the system [Esteva et al., 2001]. More sophisticated models of institutions leave to the agents' autonomy the decision whether to comply or not with the specified rules and norms of the institution [Ågotnes et al., 2007, Lopez y Lopez et al., 2004]. However, all previous models abstract away from the legislative source of the norms of an institution, and from how institutions are created, maintained and changed by their members. More precisely, while it is widely shared in the MAS field that, in order to face complex and dynamical problems, individual agents must be autonomous, less emphasis is devoted to the fact that MASs themselves for exactly the same reasons should be conceived and designed to be autonomous. In fact, etymologically, autonomous means self-binding ('auto' and 'nomos'), and an autonomous MAS should be the vision of an artificial society that is able to create, maintain, and eventually change its own institutions by itself, without the intervention of the external designer in this process.

The aim of this work is to advance the state of the art on artificial institutions and normative multi-agent systems by proposing a logical model in which the existence and the dynamics of an institution (norms, rules, institutional facts, *etc.*) are determined by the individual and collective attitudes of the agents which identify themselves as members of the institution. In particular, we propose a model in which an institution is grounded on the (individual and collective) *acceptances* of its members, and its dynamics depends on the dynamics of these acceptances. On this aspect we agree with [Mantzavinos et al., 2004], when the authors say that (p. 77):

> "only because institutions are anchored in peoples minds do they ever become behaviorally relevant. The *elucidation of the internal aspect is the crucial step* in adequately explaining the emergence, evolution, and effects of institutions." [Emphasis added].

This relationship between acceptance and institutions has been emphasized in the philosophical doctrine of Legal Positivism [Hart, 1992]. According to Hart, the foundations of a normative system or institution consist of adherence to, or acceptance of, an ultimate rule of recognition by which the validity of any rule of the institution may be evaluated. [1]

[1] In Hart's theory, the rule of recognition is the rule which specifies the ultimate criteria of validity in a legal system.

Other authors working in the field of multi-agent systems have advocated the need for a bottom up approach to the explanation of the origin and the evolution of institutions. According to these authors, institutions and their dynamics should be anchored in the agents' attitudes [Conte et al., 1998, Boella and van der Torre, 2007]. For instance, in agreement with Hart's theory, [Conte et al., 1998] have stressed that the existence of a norm in an institution (but also in a group, organization, *etc.*) depends on the recognition and acceptance of the norm by the members of the institution. In their perspective, agents in a multi-agent system contribute to the enforcement and the propagation of the norm in the social context.

The fundamental concept in our paper is that of acceptance *qua* member of an institution. This notion will be informally presented in Section 2. In Section 3 we will introduce a modal logic (called $\mathcal{AL}$ for *Acceptance Logic*) which enables to reason about acceptances of agents and groups of agents. We call the former *individual acceptances*, and the latter *collective acceptances*. In Section 4 we will study the logical properties of the notion of acceptance and its interactions with classical notions such that of individual (private) belief and that of mutual belief. On the basis of the concept of acceptance *qua* member of an institution, we will specify how a group of agents can create and maintain normative and institutional facts which hold only in an attitude-dependent way. That is, it is up to the agents, and not to the external designer, to support such facts (Section 5). Then, we will distinguish regulative components and non-regulative components of an institution [Searle, 1995] (Section 6). On the one hand, we will formalize the concept of *constitutive rule*, that is, the kind of rules accepted by the members of an institution which express classifications between different concepts and establish the relations between "brute" physical facts and institutional facts within the context of the institution (Section 6.1). Since [Searle, 1995, Searle, 1969] and [Jones and Sergot, 1996], these rules have been expressed in terms of assertions of the form "X counts as Y in the context of institution x" (*e.g.* in the institutional context of US, a piece of paper with a certain shape, color, *etc.* counts as a five-dollar bill). On the other hand, *regulative rules* will be formalized through a notion of obligation and a notion of permission by studying a reduction of deontic logic to the logic of acceptance (Section 6.2). Section 7 will be devoted to show how the logic of acceptance $\mathcal{AL}$ can be appropriately refined in order to capture some essential properties of legal institutions in which a special kind of agents called *legislators* are introduced. We will discuss some general principles which seem adequate for a formal characterization of legal institutions. Finally, in Section 8, we will compare our proposal with related logical works on institutions and normative systems. Special emphasis will be devoted to the comparison between our approach and the modal logic of normative systems and "counts-as" proposed by Grossi et al. [Grossi et al., 2006]. Proofs of the main theorems presented in the paper are collected in the annex.

2 The concept of acceptance

Some conceptual clarifications of the concept of acceptance *qua* member of an institution are needed because of the crucial role it plays in explaining the maintenance of social institutions.

Several authors have emphasized the difference between acceptance and belief as particular kinds of *individual* attitudes. Whereas private beliefs have been studied for decades [Hintikka, 1962] as representative of doxastic mental states, acceptances have only been examined since [Stalnaker, 1984] and since [Cohen, 1992]. Some authors (*e.g.* [Clarke, 1994]) claim that acceptance implies belief (at least to some minimal degree as argued in [Tollefsen, 2003]). On the contrary, in [Stalnaker, 1984] acceptance is considered to be stronger than belief. Although belief and acceptance seem very close, several authors [Bratman, 1992, Cohen, 1992, Tuomela, 2000] have argued for the importance of keeping the two notions independent. We here agree with this point of view (see Section 4.3).

For the aims of this paper we are particularly interested in a particular feature of acceptance, namely the fact that acceptance is context-dependent (on this point see also [Engel, 1998]). In our approach, this feature is directly encoded in the formal definition of acceptance (see Section 3.1). In fact, one can decide (say for prudential reasons) to reason and act by "accepting" the truth of a proposition in a specific context, and reject the very same proposition in a different context. We will explore the role of acceptance in institutional contexts. Institutional contexts are conceived here 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. On the background of such contexts, we are interested in the agents' attitudes that can be formally captured. In the context of Clue, for instance, an agent accepts that something has happened *qua* player of Clue. The state of acceptance *qua* member of an institution is the kind of acceptance one is committed to when one is "functioning as a member of the institution" [Tuomela, 2002]. In these situations it may happen that the agent's acceptances are in conflict with his/her beliefs. For instance, a lawyer who is trying to defend a client in a murder case accepts *qua* lawyer that his/her client is innocent, even she/he believes the contrary.

There exist others differences between belief and acceptance that are not encoded in our formalization of acceptance. According to [Hakli, 2006], the key difference between belief and acceptance is that the former is aimed at truth, whilst the latter depends on an agent's decision. More precisely, while a belief that p is an attitude constitutively aimed at the truth of p, an acceptance is the output of "a decision to treat p as true in one's utterances and actions" without being *necessarily* (see [Tuomela, 2000] for instance) connected to the actual truth of the proposition.

In the present paper the notion of acceptance *qua* member of an institution is also applied to the collective level named *collective acceptance*. The idea of collective attitudes is developed by Searle [Searle, 1995] among others: without supposing the existence of any collective consciousness, he argues that attitudes can be ascribed to a group of agents and that "the forms of collective intentionality cannot (...) be reduced to something else" [Searle, 1995][2].

Collective attitudes such as collective acceptance have been studied in social philosophy in opposition to the traditional notions of *mutual belief* and *mutual knowl-*

[2] A deeper discussion on this point remains out of the scope of this paper. Some interesting arguments for collective intentionality can be found in [Tollefsen, 2002].

edge that are very popular in artificial intelligence and theoretical computer science [Fagin et al., 1995, Lewis, 1969]. It has been stressed that, while mutual belief is strongly linked to individual beliefs and can be reduced to them, collective attitudes such as collective acceptance cannot be reduced to a composition of individual attitudes. This aspect is particularly emphasized by Gilbert [Gilbert, 1987] who follows Durkheim's non-reductionist view of collective attitudes [Durkheim, 1982]. According to Gilbert, any proper group attitude cannot be defined only as a label on a particular configuration of individual attitudes, as mutual belief is. In [Gilbert, 1989] it is suggested that a collective acceptance of a set of agents C is based on the fact that the agents in C identify themselves as members of a certain group, institution, team, organization, *etc.* and recognize each other as members of the same group, institution, team, organization, *etc.* (this is the view that we adopt in our formalization of acceptance, see Section 3). But mutual belief (and mutual knowledge) does not entail this aspect of mutual recognition and identification with respect to the same social context.

In accordance with [Tuomela, 2002], in this paper we consider collective acceptance with respect to institutional contexts as an attitude that is held by a set of agents *qua* members of the same institution. A collective acceptance held by a set of agents C *qua* members of a certain institution x is the kind of acceptance the agents in C are committed to when they are "functioning together as members of the institution x", that is, when the agents in C identify and recognize each other as members of the institution x. For example, in the context of the institution Greenpeace agents (collectively) accept that their mission is to protect the Earth *qua* members of Greenpeace. The state of acceptance *qua* members of Greenpeace is the kind of acceptance these agents are committed to when they are functioning together as members of Greenpeace, that is, when they identify and recognize each other as members of Greenpeace.

3 Acceptance logic

The logic $\mathcal{AL}$ (*Acceptance Logic*) enables expressing that some agents identify themselves as members of a certain institution and what (groups of) agents accept while functioning together as members of an institution. The principles of $\mathcal{AL}$ clarify the relationships between individual acceptances (acceptances of individual agents) and collective acceptances (acceptances of groups of agents).

3.1 Syntax

The syntactic primitives of $\mathcal{AL}$ are the following: a finite non-empty set of agents AGT; a countable set of atomic formulas ATM; and a finite set of labels $INST$ denoting institutions. We note $2^{AGT\star} = 2^{AGT} \setminus \{\emptyset\}$ the set of all non-empty subsets of AGT. The language $\mathcal{L}_{\mathcal{AL}}$ of the logic $\mathcal{AL}$ is given by the following BNF:

$$\varphi \ ::= \ \bot \mid p \mid \neg\varphi \mid \varphi \vee \varphi \mid \mathcal{A}_{C:x}\varphi$$

where p ranges over ATM, C ranges over $2^{AGT\star}$ and x ranges over $INST$. We define $\wedge$, $\rightarrow$, $\leftrightarrow$ and $\top$ from $\vee$, $\neg$ and $\bot$ in the usual manner.

The formula $\mathcal{A}_{C:x}\varphi$ reads "the agents in C accept that φ while functioning together as members of the institution x". For notational convenience, we write $i{:}x$ instead of $\{i\}{:}x$.

For example, $\mathcal{A}_{C:Greenpeace}protectEarth$ expresses that the agents in C accept that the mission of Greenpeace is to protect the Earth while functioning together as activists in the context of Greenpeace; and $\mathcal{A}_{i:Catholic}PopeInfallibility$ expresses that agent i accepts that the Pope is infallible while functioning as a member of the Catholic Church.

The formula $\mathcal{A}_{C:x}\bot$ has to be read "agents in C are not functioning together as members of the institution x", because we assume that functioning as a member of an institution is, at least in this minimal sense, a rational activity. Conversely, $\neg\mathcal{A}_{C:x}\bot$ has to be read "agents in C are functioning together as members of the institution x". Thus, $\neg\mathcal{A}_{C:x}\bot \wedge \mathcal{A}_{C:x}\varphi$ stands for "agents in C are functioning together as members of the institution x and they accept that φ while functioning together as members of x" or simply "agents in C accept that φ *qua* members of the institution x". Therefore $\neg\mathcal{A}_{C:x}\varphi$ has to be read "agents in C do not accept that φ be true *qua* members of x".

3.2 $\mathcal{AL}$ frames

We use a standard possible worlds semantics. Let the set of all couples of non-empty subsets of agents and institutional contexts be

$$\Delta = 2^{AGT^\star} \times INST.$$

A *frame* of the logic of acceptance $\mathcal{AL}$ ($\mathcal{AL}$ *frame*) is a couple

$$\mathcal{F} = \langle W, \mathscr{A} \rangle$$

where:

- W is a non-empty set of possible worlds;
- $\mathscr{A} : \Delta \to W \times W$ maps every $C{:}x \in \Delta$ to a relation $\mathscr{A}_{C:x}$ between possible worlds in W.

We note $\mathscr{A}_{C:x}(w) = \{w' : \langle w, w' \rangle \in \mathscr{A}_{C:x}\}$ the set of worlds that the agents in C accept at w while functioning together as members of the institution x.

We impose the following constraints on $\mathcal{AL}$ frames, for any world $w \in W$, institutional context $x \in INST$, and sets of agents $C, B \in 2^{AGT^\star}$ such that $B \subseteq C$:

(S.1) if $w' \in \mathscr{A}_{B:y}(w)$ then $\mathscr{A}_{C:x}(w') \subseteq \mathscr{A}_{C:x}(w)$

(S.2) if $w' \in \mathscr{A}_{B:y}(w)$ then $\mathscr{A}_{C:x}(w) \subseteq \mathscr{A}_{C:x}(w')$

(S.3) if $\mathscr{A}_{C:x}(w) \neq \emptyset$ then $\mathscr{A}_{B:x}(w) \subseteq \mathscr{A}_{C:x}(w)$

(S.4) if $w' \in \mathscr{A}_{C:x}(w)$ then $w' \in \bigcup_{i \in C} \mathscr{A}_{i:x}(w')$

(S.5) if $\mathscr{A}_{C:x}(w) \neq \emptyset$ then $\mathscr{A}_{B:x}(w) \neq \emptyset$

The constraint **S.1** is a generalized version of transitivity: given two sets of agents C, B such that $B \subseteq C$, if w' is a world that the agents in B accept at w while functioning together as members of the institution y and w'' is a world that the agents in C

accept at w' while functioning together as members of the institution x then, w'' is a world that the agents in C accept at w while functioning together as members of the institution x.

The constraint **S.2** is a generalized version of euclideanity: given two sets of agents C, B such that $B \subseteq C$, if w' is a world that the agents in B accept at w while functioning together as members of the institution y and w'' is a world that the agents in C accept at w while functioning together as members of the institution x then, w'' is a world that the agents in C accept at w' while functioning together as members of the institution x.

The constraint **S.3** is a property of conditional inclusion: given two sets of agents C, B such that $B \subseteq C$, if there exists a world w'' that the agents in C accept at w while functioning together as members of the institution x and w' is a world that the agents in B accept at w while functioning together as members of the institution x then, w' is also a world that the agents in C accept at w while functioning together as members of the institution x.

The constraint **S.4** is a sort of weak reflexivity: if w' is a world that the agents in C accept at w while functioning together as members of the institution x then, there exists some agent $i \in C$ such that w' is a world that agent i accepts at w', while functioning as a member of the institution x.

According to the last constraint **S.5**, given two sets of agents C, B such that $B \subseteq C$, if there exists a world w' that the agents in C accept at w while functioning together as members of the institution x then, there exists a world w'' that the agents in B accept at w while functioning together as members of the institution x.

3.3 $\mathcal{AL}$ models and validity

A *model* of the logic of acceptance $\mathcal{AL}$ ($\mathcal{AL}$ *model*) is a couple

$$\mathcal{M} = \langle \mathcal{F}, \mathscr{V} \rangle$$

where:

- $\mathcal{F}$ is a $\mathcal{AL}$ frame;
- $\mathscr{V} : ATM \rightarrow 2^W$ is valuation function associating a set of possible worlds $\mathscr{V}(p) \subseteq W$ to each atomic formula p of ATM.

Given $\mathcal{M} = \langle W, \mathscr{A}, \mathscr{V} \rangle$ and $w \in W$, the couple $\langle \mathcal{M}, w \rangle$ is a *pointed* $\mathcal{AL}$ *model*. Given a formula φ, we write $\mathcal{M}, w \models \varphi$ and say that φ is *true* at world w in $\mathcal{M}$. The notation $\mathcal{M}, w \not\models \varphi$ means that φ is *false* at world w in $\mathcal{M}$. The truth conditions for the formulas of the logic $\mathcal{AL}$ are:

- $\mathcal{M}, w \not\models \bot$;
- $\mathcal{M}, w \models p$ iff $w \in \mathscr{V}(p)$;
- $\mathcal{M}, w \models \neg\varphi$ iff $\mathcal{M}, w \not\models \varphi$;
- $\mathcal{M}, w \models \varphi \vee \psi$ iff $\mathcal{M}, w \models \varphi$ or $\mathcal{M}, w \models \psi$;
- $\mathcal{M}, w \models \mathcal{A}_{C:x}\varphi$ iff $\mathcal{M}, w' \models \varphi$ for all $w' \in \mathscr{A}_{C:x}(w)$.

A formula φ is *true in a $\mathcal{AL}$ model* $\mathcal{M}$ if and only if $\mathcal{M}, w \models \varphi$ for every world w in $\mathcal{M}$. φ is $\mathcal{AL}$ *valid* (noted $\models_{\mathcal{AL}} \varphi$) if and only if φ is true in all $\mathcal{AL}$ models. φ is $\mathcal{AL}$ *satisfiable* if and only if $\neg\varphi$ is not $\mathcal{AL}$ valid.

3.4 Axiomatization

The axiomatization of $\mathcal{AL}$ is as follows:

(**ProTau**)	All principles of propositional calculus
(**K**)	$\mathcal{A}_{C:x}(\varphi \to \psi) \to (\mathcal{A}_{C:x}\varphi \to \mathcal{A}_{C:x}\psi)$
(**PAccess**)	$\mathcal{A}_{C:x}\varphi \to \mathcal{A}_{B:y}\mathcal{A}_{C:x}\varphi$ if $B \subseteq C$
(**NAccess**)	$\neg\mathcal{A}_{C:x}\varphi \to \mathcal{A}_{B:y}\neg\mathcal{A}_{C:x}\varphi$ if $B \subseteq C$
(**Inc**)	$(\neg\mathcal{A}_{C:x}\bot \wedge \mathcal{A}_{C:x}\varphi) \to \mathcal{A}_{B:x}\varphi$ if $B \subseteq C$
(**Unanim**)	$\mathcal{A}_{C:x}(\bigwedge_{i \in C} \mathcal{A}_{i:x}\varphi \to \varphi)$
(**Mon**)	$\neg\mathcal{A}_{C:x}\bot \to \neg\mathcal{A}_{B:x}\bot$ if $B \subseteq C$
(**MP**)	From $\vdash \varphi$ and $\vdash \varphi \to \psi$ infer $\vdash \psi$
(**Nec**)	From $\vdash \varphi$ infer $\vdash \mathcal{A}_{C:x}\psi$

This axiomatization includes all tautologies of propositional calculus (**ProTau**) and the rule of inference *modus ponens* (**MP**). Axiom **K** and rule of *necessitation* (**Nec**) define a minimal normal modal logic. (See [Chellas, 1980, chap. 4].)

Axioms **PAccess** and **NAccess** express that a group of agents has always access to what is accepted (resp. not accepted) by its supergroups.

Axiom **PAccess** concerns the (positive) access to what is accepted by a supergroup: when the agents in a set C function together as members of the institution x, then for all $B \subseteq C$ the agents in B have access to all facts that are accepted by the agents in C. That is, if the agents in C accept that φ while functioning together as members of the institution x then, while functioning together as members of x, the agents of every subset B of C accept that the agents in C accept that φ.

Axiom **NAccess** concerns the (negative) access to what is not accepted by a supergroup: if the agents in C do not accept that φ while functioning together as members of the institution x then, while functioning together as members of x, the agents of every subset B of C accept that the agents in C do not accept that φ.

Example 1. *Suppose that three agents i, j, k, while functioning together as members of the UK trade union, accept that their mission is to increase teachers' wages, but they do not accept* qua *members of the trade union that their mission is to increase railway workers' wages:*

$\mathcal{A}_{\{i,j,k\}:Union}\, increaseTeacherWage$ *and* $\neg\mathcal{A}_{\{i,j,k\}:Union}\, increaseRailwayWage$.

By Axiom ***PAccess*** *we infer that, while functioning as a UK citizen, i accepts that i, j, k accept that their mission is to increase teachers' wages, while functioning together as members of the trade union:*

$$\mathcal{A}_{i:UK}\mathcal{A}_{\{i,j,k\}:Union}\, increaseTeacherWage.$$

By Axiom ***NAccess*** *we infer that, while functioning as a UK citizen, i accepts that i, j, k do not accept,* qua *members of the trade union, that their mission is to increase railway workers' wages:*

$$\mathcal{A}_{i:UK} \neg \mathcal{A}_{\{i,j,k\}:Union}\, increaseRailwayWage.$$

Axiom **Inc** says that, if the agents in C accept that φ *qua* members of x then for every subset B of C the agents in B accept φ while functioning together as members of x. This means that the facts accepted by the agents in C *qua* members of a certain institution x are necessarily accepted by the agents in all of C's subsets with respect to the same institution. Therefore Axiom **Inc** describes the *top down* process leading from C's collective acceptance to the individual acceptances of the agents in C.

Example 2. *Imagine three agents i, j, k that,* qua *players of the game Clue, accept that someone called Mrs. Red, has been killed:*

$$\neg \mathcal{A}_{\{i,j,k\}:Clue} \bot \wedge \mathcal{A}_{\{i,j,k\}:Clue} killedMrsRed.$$

By Axiom ***Inc*** *we infer that also the two agents i, j, while functioning as Clue players, accept that someone called Mrs. Red has been killed:*

$$\mathcal{A}_{\{i,j\}:Clue} killedMrsRed.$$

Axiom **Unanim** expresses a unanimity principle according to which the agents in C, while functioning together as members of x, accept that if each of them individually accepts that φ while functioning as a member of x, then φ is the case. This axiom describes the *bottom up* process leading from the individual acceptances of the members of C to the collective acceptance of the group C.

Finally, Axiom **Mon** expresses an intuitive property of monotonicity about institution membership. It says that, if the agents in C are functioning together as members of the institution x then, for every subset B of C, the agents in B are also functioning together as members of the institution x. As emphasized in Section 2, "the agents in C function together as members of institution x" means for us that "the agents in C identify and recognize each other as members of the same institution x". Thus, Axiom **Mon** can be rephrased as follows: if the agents in a set C identify and recognize each other as members of the institution x then, for every subset B of C, the agents in B also identify and recognize each other as members of x.

The following correspondences (in the sense of correspondence theory, see for instance [van Benthem, 2001, Blackburn et al., 2001]) exist between the axioms of the logic $\mathcal{AL}$ and the semantic constraints over $\mathcal{AL}$ frames given in Section 3.2 (see also proof of Theorem 1 in the Annex): Axiom **PAccess** corresponds to the constraint **S.1**, **NAccess** corresponds to **S.2**, **Inc** corresponds to **S.3**, **Unanim** corresponds to **S.4** and **Mon** corresponds to **S.5**.

We call $\mathcal{AL}$ the logic axiomatized by the principles given above: **ProTau**, **K**, **PAccess**, **NAccess**, **Inc**, **Unanim**, **Mon**, **MP**, **Nec**. We write $\vdash_{\mathcal{AL}} \varphi$ if formula φ is a theorem of $\mathcal{AL}$ and $\nvdash_{\mathcal{AL}} \varphi$ if formula φ is not a theorem.

We can prove that $\mathcal{AL}$ is sound and complete with respect to the class of $\mathcal{AL}$ frames.

Theorem 1. $\vdash_{\mathcal{AL}} \varphi$ *if and only if* $\models_{\mathcal{AL}} \varphi$.

By the standard filtration method we can also prove that the logic $\mathcal{AL}$ is decidable.

Theorem 2. *The logic $\mathcal{AL}$ is decidable.*

In the following section the properties of the concepts of individual acceptance, collective acceptance and institution membership will be studied. We will also study the relationships between acceptance and belief in a more formal way than in Section 2.

4 General properties

4.1 Properties of acceptance and institution membership

The following theorem highlights some interesting properties of collective acceptance and institution membership.

Theorem 3. *For every $x, y \in INST$ and $B, C \in 2^{AGT^\star}$ such that $B \subseteq C$:*

(3a) $\vdash_{\mathcal{AL}} \mathcal{A}_{C:x} \neg \mathcal{A}_{C:x} \bot$

(3b) $\vdash_{\mathcal{AL}} \mathcal{A}_{C:x} \bigwedge_{i \in C} \neg \mathcal{A}_{i:x} \bot$

(3c) $\vdash_{\mathcal{AL}} \mathcal{A}_{B:y} \mathcal{A}_{C:x} \varphi \leftrightarrow \mathcal{A}_{C:x} \varphi$

(3d) $\vdash_{\mathcal{AL}} \mathcal{A}_{B:y} \neg \mathcal{A}_{C:x} \varphi \leftrightarrow (\mathcal{A}_{B:y} \bot \vee \neg \mathcal{A}_{C:x} \varphi)$

(3e) $\vdash_{\mathcal{AL}} \mathcal{A}_{C:x} (\mathcal{A}_{C:x} \varphi \rightarrow \varphi)$

(3f) $\vdash_{\mathcal{AL}} (\mathcal{A}_{C:x} \bigwedge_{i \in C} \mathcal{A}_{i:x} \varphi) \leftrightarrow \mathcal{A}_{C:x} \varphi$

Theorem 3a expresses a property of institution membership. It says that the agents in a group C, while functioning together as members of the institution x, accept that they are functioning together as members of the institution x. Theorem 3b is another way to express the property of institution membership: it expresses that the agents in a group, while functioning together as members of a certain institution, accept that everyone of them is functioning as a member of the institution.

Example 3. *Suppose that, during a concert, the agents in C are functioning together as members of the Philharmonic Orchestra. Then, according to Theorem 3a, this fact is accepted by the group C. That is, while functioning together as members of the Philharmonic Orchestra, the agents in C accept that they are functioning together as members of the Philharmonic Orchestra: $\mathcal{A}_{C:Orchestra} \neg \mathcal{A}_{C:Orchestra} \bot$. Moreover, they accept that everyone of them is functioning as a member of the Philharmonic Orchestra: $\mathcal{A}_{C:Orchestra} \bigwedge_{i \in C} \neg \mathcal{A}_{i:Orchestra} \bot$.*

Theorem 3c and Theorem 3d together express that a group of agents B can never be wrong in ascribing a collective acceptance to its supergroup C and in recognizing that its supergroup C does not accept something. Furthermore, a group of agents B has always correct access to what is accepted (resp. not accepted) by its supergroups. The right to left direction of Theorem 3c is Axiom **PAccess**. The left to right direction means that, given two sets of agents B and C such that $B \subseteq C$, if the agents in B, while functioning together as members of institution y, accept that the agents in C accept φ while functioning together as members of institution x then, the agents in C accept φ

while functioning together as members of institution x. The right to left direction of Theorem 3d is Axiom **NAccess**. The left to right direction means that, given two sets of agents B and C such that $B \subseteq C$, if the agents in B, while functioning together as members of institution y, accept that the agents in C do not accept φ *qua* members of institution x then, either the agents in B do not function as members of y or the agents in C do not accept φ *qua* members of institution x.

Theorem 3e and Theorem 3f are variants of the unanimity Axiom **Unanim**. Theorem 3e says that for every set of agents C, the agents in C, while functioning together as members of x, accept that if they accept that φ while functioning together as members of x, then φ is the case. Theorem 3f expresses that: if the agents in C, while functioning together as members of x, accept that each of them individually accepts that φ while functioning as a member of x, then the agents in C, while functioning together as members of x, accept that φ is the case.

The following theorem highlights the relationship between the acceptance of a group of agents and the acceptances of its subgroups.

Theorem 4. *For every $x \in INST$ and $C_1, C_2, C_3 \in 2^{AGT}$ such that $C_3 \subseteq C_2 \subseteq C_1$ and $C_3 \neq \emptyset$:*

$$\vdash_{\mathcal{AL}} \mathcal{A}_{C_1:x}(\mathcal{A}_{C_2:x}\varphi \rightarrow \mathcal{A}_{C_3:x}\varphi)$$

Theorem 4 expresses that every group of agents has to accept the principle of inclusion formalized by Axiom **Inc**.

4.2 Discussion around the unanimity principle

Let us consider more in detail the unanimity property of our logic of acceptance expressed by Axiom **Unanim** (and Theorems 3e,3f). This property says that collective acceptances emerge from consensus. This is for us a necessary requirement for a notion of collective acceptance which is valid for all institutions and groups. We did not include stronger principles which explain how a collective acceptance of a group of agents C might be constructed. Nevertheless, one might go further and consider other kinds of principles which are specific to certain institutions and groups.

For example, one might want to extend the analysis to formal (legal) institutions in which special agents with the power to affect the acceptances of the other members of the institution are introduced. In legal institutions, one can formalize the rule according to which all facts that are accepted by the legislators of an institution must be universally accepted by all members of this institution. Suppose that x denotes a legal institution (*e.g.* EU, Association of Symbolic Logic, *etc.*) which has a non-empty set of agents called legislators, noted $Leg(x) \in 2^{AGT^\star}$. (See Section 7 for a precise definition of the function $Leg()$ and a more elaborate analysis of the concepts of legislator and legal institution.) From this, one can formalize a principle stating that everything that the legislators of the legal institution x accept is universally accepted in the legal institution x:

(Legislators) $$\mathcal{A}_{C:x}(\bigwedge_{i \in Leg(x)} \mathcal{A}_{i:x}\varphi \rightarrow \varphi)$$

The Principle **Legislators** says that, for every group of agents C, while functioning together as members of the institution x, the agents in C accept that if the legislators of x accept that φ, then φ is the case.

Another interesting principle for the construction of collective acceptance is majority. (In this case, unanimity is not required to obtain a consensus.) This kind of principle applies both to informal and formal institutions. The principle of majority could be introduced as a logical axiom for two specific sets of agents C and B such that $B \subseteq C$ and $|C \setminus B| < |B|$ (*i.e.* B represents the majority of agents in C):

(**Majority**) $$\mathcal{A}_{C:x}(\bigwedge_{i \in B} \mathcal{A}_{i:x}\varphi \rightarrow \varphi)$$

The Principle **Majority** says that, for every group of agents C, while functioning together as members of the institution x, the agents in C accept that if the majority of them accept that φ, then φ is the case. The following example by Pettit [Pettit, 2001] shows how the majority principle would work.

Example 4. *Imagine a three-member court which has to make a judgment on whether a defendant is liable (noted l) for a breach of contract. The three judges i, j and k accept a majority rule to decide on the issue. That is, i, j and k, while functioning together as members of the court, accept that if the majority of them accepts that the defendant is liable (resp. not liable), then the defendant is liable (resp. not liable). Formally, for any B such that $B \subseteq \{i, j, k\}$ and $|B| = 2$ we have:*

$$\mathcal{A}_{\{i,j,k\}:court}(\bigwedge_{i \in B} \mathcal{A}_{i:court}l \rightarrow l) \wedge \mathcal{A}_{\{i,j,k\}:court}(\bigwedge_{i \in B} \mathcal{A}_{i:court}\neg l \rightarrow \neg l)$$

Therefore, if the three judges accept that two of them accept that the defendant is liable, i.e. $\mathcal{A}_{\{i,j,k\}:court}(\mathcal{A}_{i:court}l \wedge \mathcal{A}_{j:court}l)$*, by the Principle **Majority** and Axiom **K** it follows that the three judges have to accept that the judge is liable,* i.e. $\mathcal{A}_{\{i,j,k\}:court}l$*).*

It has to be noted that the previous principle of majority cannot be generalized to all sets of agents without incurring the following very counterintuitive consequence.

Proposition 1. *If we suppose that the Principle **Majority** is valid for any B, C such that $B \subseteq C$ and $|C \setminus B| < |B|$ then, the following consequence is derivable, for $i \neq j$:*

$$(\mathcal{A}_{AGT:x}\mathcal{A}_{\{i,j\}:x}\varphi \wedge \neg\mathcal{A}_{AGT:x}\bot) \rightarrow \mathcal{A}_{AGT:x}\varphi$$

This means that, when the majority principle is generalized to all sets of agents, we can infer that: if all agents, *qua* members of institution x, accept that two of them accept φ while functioning together as members of institution x then, the acceptances of the two agents propagate to all agents in such a way that all agents accept φ *qua* members of institution x.

4.3 Relationships between acceptance and belief

As said in Section 2, there is a large literature about the distinction between belief and acceptance. For us, belief and acceptance are clearly different concepts in several

senses. In this section we focus on the distinction between acceptance, individual belief and mutual belief. Our aim is to provide further clarifications of the concept of acceptance in terms of its relationships with other kinds of agents' attitudes rather than proposing an extension of the logic $\mathcal{AL}$ with individual belief and mutual belief and studying its mathematical properties. Here, we just show how modal operators for belief and mutual belief can be integrated into the logic $\mathcal{AL}$ on the basis of some intuitive interaction principles relating acceptance and belief.

For convenience, we note $Bel_i\varphi$ the formula that reads "the agent i believes that φ is true", and we suppose that belief operators of type Bel_i are defined as usual in a KD45 modal logic [Hintikka, 1962]. Belief operators Bel_i are interpreted in terms of accessibility relations $\mathscr{B}_i$ on the set of possible worlds W. These accessibility relations are supposed to be serial, transitive and euclidean. We write $\mathscr{B}_i(w)$ for the set $\{w' : \langle w, w'\rangle \in \mathscr{B}_i\}$. $\mathscr{B}_i(w)$ is the set of worlds that are possible according to agent i. The truth condition is:

$$\mathcal{M}, w \models Bel_i\varphi \quad \text{iff} \quad \mathcal{M}, w' \models \varphi \text{ for every } w' \in \mathscr{B}_i(w)$$

Moreover we introduce the notion of mutual belief which has been extensively studied both in the computer science literature [Fagin et al., 1995] and in the philosophical literature [Lewis, 1969]. Given a set of agents $C \subseteq AGT$, $\mathcal{MB}_C\varphi$ reads "there is a mutual belief in C that φ", that is, "everyone in C believes that φ, everyone in C believes that everyone in C believes that φ, everyone in C believes that everyone in C believes that everyone in C believes that φ, and so on". The mutual belief of a set of agents C is interpreted in terms of the transitive closure $\mathscr{B}_C^+$ of the union of the accessibility relations $\mathscr{B}_i$ for every agent $i \in C$, that is:

$$\mathcal{M}, w \models \mathcal{MB}_C\varphi \quad \text{iff} \quad \mathcal{M}, w' \models \varphi \text{ for every } w' \in \mathscr{B}_C^+(w)$$

Let the concept of "everybody in group C believes φ" be defined as follows:

$$E_C\varphi \stackrel{def}{=} \bigwedge_{i \in C} Bel_i\varphi$$

As shown in [Fagin et al., 1995], the following axioms and rules of inference provide a sound and complete axiomatization of the logic of individual belief and mutual belief:

(**KD45$_{\mathbf{Bel}}$**)	All KD45-principles for the operators Bel_i
(**FixPoint**)	$\vdash \mathcal{MB}_C\varphi \leftrightarrow E_C(\varphi \wedge \mathcal{MB}_C\varphi)$
(**InductionRule**)	From $\vdash \varphi \rightarrow E_C(\varphi \wedge \psi)$ infer $\vdash \varphi \rightarrow \mathcal{MB}_C\psi$

The first interesting thing to note is that, although collective acceptance and mutual belief have different natures (see the discussion in Section 2), they share the Fix Point property. The following Theorem 5 highlights this aspect.

Theorem 5.

$$\vdash_{\mathcal{AL}} \mathcal{A}_{C:x}\varphi \leftrightarrow \bigwedge_{i \in C} \mathcal{A}_{i:x}(\varphi \wedge \mathcal{A}_{C:x}\varphi)$$

Nevertheless we cannot argue that our concept of collective acceptance is stronger than the concept of mutual belief, in particular because the **InductionRule** does not hold in $\mathcal{AL}$. This is due to the non-reductionist feature of the collective acceptance: it cannot be reduced to a particular configuration of individual acceptances.

The following two sections are devoted to discuss other interesting relations between acceptance and belief. We will first provide an analysis of the shared aspect of collective acceptance expressed in terms of mutual belief. Then, we will briefly consider the problem of the incompatibility between acceptance and belief.

4.3.1 The shared nature of collective acceptance

As emphasized in the philosophical literature [Gilbert, 1989, Tuomela, 1992], a collective acceptance of the agents in a set C must not be confused with (nor reduced to) the sum of the individual acceptance of the agents in C. On the contrary, when the agents in C accept some fact φ to be true *qua* members of a certain institution, it means that every agent in C declares to the other agents of the group C that she/he is willing to accept φ to be true. This aspect of acceptance can be formally derived by supposing the following two principles relating individual beliefs with collective acceptances.

(**PIntrAccept**)	$\mathcal{A}_{C:x}\varphi \rightarrow Bel_i\mathcal{A}_{C:x}\varphi$	if $i \in C$
(**NegIntrAccept**)	$\neg\mathcal{A}_{C:x}\varphi \rightarrow Bel_i\neg\mathcal{A}_{C:x}\varphi$	if $i \in C$

The first principle says that: if the agents in C accept that φ while functioning together as members of the institution x then, every agent in C believes this. The second principle says that: if the agents in C do not accept φ *qua* members of x then every agent in C believes this.

We can easily prove that, under the previous two principles, collective acceptance is always shared so much that the group C accepts φ if and only if the agents in C mutually believe this. More formally:

Proposition 2. *For any $C{:}x \in \Delta$, the following formulas are derivable from the axiom **D** for belief (following from **KD45**$_{\mathbf{Bel}}$), Axiom **FixPoint** and Rule of inference **InductionRule** for mutual belief, and the interaction Principles **PIntrAccept** and **NegIntrAccept** for acceptance and belief.*

(2a)	$\mathcal{A}_{C:x}\varphi \leftrightarrow \mathcal{MB}_C\mathcal{A}_{C:x}\varphi$
(2b)	$\neg\mathcal{A}_{C:x}\varphi \leftrightarrow \mathcal{MB}_C\neg\mathcal{A}_{C:x}\varphi$

According to Proposition 2a, the agents in C accept that φ while functioning together as members of the institution x if and only if there is a mutual belief in C that they accept that φ while functioning together as members of the institution x. According to Proposition 2b, the agents in C do not accept that φ *qua* members of x if and only if there is a mutual belief in C that they do not accept that φ *qua* members of x. Hence, accepting (resp. not accepting) a proposition while functioning as members of an institution is always a *mutually believed* fact (for the members of the group) which is out in the open and that is used by all the members to reason about each other in the institutional context.

4.3.2 Acceptance and belief might be incompatible

Individual belief and individual acceptance are both private mental attitudes but: an individual belief does not depend on context, whilst an individual acceptance is a context-dependent attitude which is entertained by an agent *qua* member of a given institution. Therefore, an agent can privately disbelieve something she/he accepts while functioning as a member of a given institution. Formally: $Bel_i\varphi \wedge \mathcal{A}_{i:x}\neg\varphi$ may be true. In a similar way, as emphasized in [Tuomela, 1992], a collective acceptance that φ by a group of agents C (*qua* members of a given institution) might be compatible with the fact that none of the agents in C believes that φ (and even that every agent in C believes that $\neg\varphi$). The following example, inspired by [Tuomela, 1992, p. 285], illustrates this point.

Example 5. *At the end of the 80s, the Communist Party of Ruritania accepted that capitalist countries will soon perish (but none of its members really believed so).*

This means that the agents in C accept that capitalist countries will perish ($ccwp$) *qua* members of the Communist Party of Ruritania (CPR) but nobody in C (privately) believes this. Thus, formally: $\neg\mathcal{A}_{C:CPR}\bot \wedge \mathcal{A}_{C:CPR}ccwp \wedge \bigwedge_{i\in C}\neg Bel_i ccwp$.

In the following Section 5 we will show how institutional facts can be grounded on agents' acceptances in such a way that the existence of the former depends on the latter.

5 Truth in an institutional context

Recent theories of institutions [Lagerspetz, 2006, Searle, 1995, Tuomela, 2002] share at least the following two theses.

Performativity: the acceptance that a certain fact is true shared by the members of a certain institution may contribute to the truth of this fact within the context of the institution.

Reflexivity: if a certain fact is true within the context of a certain institution, the acceptance of this fact by the members of the institution is present.

More precisely, a certain fact φ is true within the context of an institution x if and only if the fact φ is accepted to be true by the members of the institution x. Therefore, a *necessary* condition for the existence of a fact within the context of an institution is that this fact is accepted to exist by the members of the institution. Moreover, the acceptance of a certain fact by the members of an institution is a *sufficient* condition for the existence of this fact within the context of the institution.

Example 6. *If the agents,* qua *European citizens, accept a certain piece of paper with a certain shape, color,* etc. *as money, then, within the context of EU, this piece of paper is money (performativity). At the same time, if it is true that a certain piece of paper is money within the context of EU, then the agents* qua *European citizens accept the piece of paper as money (reflexivity).*

Our aim here is to represent in $\mathcal{AL}$ those facts that are true within the context of an institution, that is, to define the concept of truth with respect to an institutional context (*institutional truth*) in a way that respects the previous two principles of reflexivity and performativity. We formalize the notion of institutional truth by means of the operator $[x]$. A formula $[x]\,\varphi$ is read "within the institutional context x, it is the case that φ". We take the latter to be synonymous of "for every set of agents C, the agents in C accept that φ while functioning together as members of the institution x". Formally, for every $x \in INST$:

$$[x]\,\varphi \stackrel{def}{=} \bigwedge_{C \in 2^{AGT^\star}} \mathcal{A}_{C:x}\varphi$$

According to our definition, a fact φ is true within the context of institution x if and only if, for every group C, the agents in C accept φ, while functioning together as members of x. Hence the performativity and the reflexivity principles mentioned above are guaranteed.

It is worth noting that this formal definition of truth with respect to an institution is perfectly adequate to characterize informal institutions in which there are no specialized agents called legislators empowered to change the institution itself on behalf of everybody else. It is a peculiar property of informal institutions the fact that they are based on the general consensus of all their members [Coleman, 1990], that is, a certain fact φ is true within the context of an informal institution x if and only if all members of x accept φ to be true. In Section 7 we will show how the operator $[x]$ can be appropriately redefined in order to characterize formal (legal) institution and to distinguish them from informal institutions. For the moment, we just suppose that our model only applies to the basic informal institutions of a society in which no legislator is given.

It is straightforward to prove that $[x]$ is a normal modal operator satisfying Axiom K and the necessitation rule.

Theorem 6. *For every $x \in INST$:*

(6a) $\qquad \vdash_{\mathcal{AL}} \quad [x]\,(\varphi \to \psi) \to ([x]\,\varphi \to [x]\,\psi)$

(6b) $\qquad$ *From* $\vdash_{\mathcal{AL}} \varphi$ *infer* $\vdash_{\mathcal{AL}} [x]\,\varphi$

Nevertheless, institutional operators of type $[x]$ fail to satisfy Axiom 4 and Axiom 5. That is, $[x]\,\varphi \wedge \neg\,[x]\,[x]\,\varphi$ and $\neg\,[x]\,\varphi \wedge \neg\,[x]\,\neg\,[x]\,\varphi$ are satisfiable in the logic $\mathcal{AL}$ for any $x \in INST$. This means that for every institution x, the members of x might accept φ while they do not accept that they accept φ and, it might be the case that the members of x do not accept φ, while they do not accept that they do not accept φ. The operator $[x]$ does to satisfy these two properties because of the restriction imposed on Axioms **PAccess** and **NAccess** according to which, the agents in a group B have access to all facts accepted (resp. not accepted) by the agents in another group C, *only if* B is a subgroup of C. Therefore, in the logic $\mathcal{AL}$, a certain fact φ might be accepted by all groups of members of a certain institution x, while some group of members of x does not have access to the fact that all groups of members of x accept φ. (See Section 8.1 for a discussion about a different point of view.)

The following operator $[Univ]$ is defined in order to express facts which are true in all institutions:

$$[Univ]\,\varphi \stackrel{def}{=} \bigwedge_{x \in INST} [x]\,\varphi$$

where $[Univ]\,\varphi$ is meant to stand for "φ is universally accepted as true". The operator $[Univ]$ is also a normal modal operator satisfying Axiom K and the necessitation rule:

Theorem 7.

(7a) $\quad \vdash_{\mathcal{AL}} \quad [Univ]\,(\varphi \to \psi) \to ([Univ]\,\varphi \to [Univ]\,\psi)$

(7b) $\quad$ *From* $\vdash_{\mathcal{AL}} \varphi$ *infer* $\vdash_{\mathcal{AL}} [Univ]\,\varphi$

The operator $[Univ]$ too fails to satisfy Axiom 4 and Axiom 5. Indeed, $[Univ]\,\varphi \wedge \neg\,[Univ]\,[Univ]\,\varphi$ and $\neg\,[Univ]\,\varphi \wedge \neg\,[Univ]\,\neg\,[Univ]\,\varphi$ are satisfiable in the logic $\mathcal{AL}$. This means that: φ might be universally accepted, while it is not universally accepted that φ is universally accepted and; it might be the case that φ is not universally accepted, while it is not universally accepted that φ is not universally accepted.

In the following section operators of institutional truth of type $[x]$ and the operator of universal truth $[Univ]$ will be used to define the concepts of *constitutive rule* and *regulative rule*. These two concepts are indeed fundamental for a theory of institutions.

6 Constitutive rules and regulative rules

According to many philosophers [Rawls, 1955, Alchourrón and Bulygin, 1971] working on social theory and researchers in the field of normative multi-agent systems [Boella and van der Torre, 2004], institutions are based both on regulative and non-regulative components. In particular, institutions are not only defined in terms of sets of permissions, obligations, and prohibitions (*i.e. norms of conduct* [Bulygin, 1992]) but also in terms of rules which specify and create new forms of behavior and concepts. Several terms such as *constitutive rule* [Searle, 1969, Searle, 1995], *conceptual rule* [Bulygin, 1992] or *determinative rule* [Von Wright, 1963] have been used to identify this non-regulative dimension of institutions. According to Searle for instance "(...) regulative rules regulate antecedently or independently existing forms of behavior (...). But constitutive rules do not merely regulate, they create or define new forms of behavior" [Searle, 1969, p. 33]. In Searle's theory of institutions [Searle, 1969, Searle, 1995], constitutive (*i.e.* non-regulative) rules are expressed by means of "counts-as" statements of the form "X counts as Y in context x" where the context x refers to the institution/normative system in which the rule is specified. As emphasized in [Grossi et al., 2006], "counts-as" statements are used to express classifications and subsumption relations between different concepts, that is, they assert just that a concept X is a subconcept of a concept Y. These classifications are fundamental for establishing the relations between "brute" physical facts and objects on the one hand, and institutional facts and objects on the other hand (*e.g.* money, private property, *etc.*). For example, in the institutional context of Europe, a piece of paper with a certain shape, color, *etc.* (a physical object) counts as a five-euro bill (an institutional object).

6.1 Constitutive rules

From the concept of institutional truth presented above, a notion of constitutive rule of the form "φ counts as ψ in the institutional context x" can be defined in the logic $\mathcal{AL}$. We conceive a constitutive rule as a material implication of the form $\varphi \rightarrow \psi$ in the scope of an operator $[x]$. Thus, "φ counts as ψ in the institutional context x" only if every group of members of institution x accepts that φ entails ψ. Furthermore, we suppose that a constitutive rule is *intrinsically contextual*, which means that the rule is not universally valid while it is accepted by the members of a certain institution. More precisely, we exclude situations in which $[Univ](\varphi \rightarrow \psi)$ is true (*i.e.* situations in which it is universally accepted that φ entails ψ).

In this perspective, "counts-as" statements with respect to a certain institutional context x do not just express that the members of institution x classify φ as ψ in virtue of their acceptances, but also that this classification is proper to the institution, *i.e.* it is not universally accepted that φ entails ψ. (See [Grossi et al., 2006] for a similar perspective.) In this sense, the notion of "counts-as" presented here is aimed at capturing the proper meaning of the term "constitutive rule", that is, a rule which constitutes something new within the context of an institution.

Thus, for every $x \in INST$ the following abbreviation is given:

$$\varphi \overset{x}{\rhd} \psi \overset{def}{=} [x](\varphi \rightarrow \psi) \wedge \neg [Univ](\varphi \rightarrow \psi)$$

where $\varphi \overset{x}{\rhd} \psi$ stands for "φ counts as ψ in the institutional context x".

Example 7. *Let consider the institutional context of gestural language. There exists a constitutive rule in this language according to which, the nodding gesture counts as an endorsement of what the speaker is suggesting,* i.e. $nodding \overset{gesture}{\rhd} yes$. *This means that every group of speakers using gestural language accepts that making the nodding gesture entails endorsing what the speaker is suggesting,* i.e. $[gesture](nodding \rightarrow yes)$*, and there are members of other institutions (*e.g. *different cultural contexts in which the same gesture does not express the same fact) who do not accept this,* i.e. $\neg [Univ](nodding \rightarrow yes)$.

Note that a stronger version of the concept of constitutive rule could be given by supposing that "φ counts as ψ in the institutional context x" if and only if φ entails ψ within the institutional context x, *i.e.* $[x](\varphi \rightarrow \psi)$, and for every institution y, if $y \neq x$ then it is not the case that φ entails ψ within the institutional context y, *i.e.* $\bigwedge_{y \in INST, y \neq x} \neg [y](\varphi \rightarrow \psi)$. The latter condition implies the condition $\neg [Univ](\varphi \rightarrow \psi)$ in the definition of the "counts-as" conditional $\varphi \overset{x}{\rhd} \psi$. This stronger version of the concept of constitutive rule is not analyzed in the present paper.

The following two theorems highlight some valid and invalid properties of "counts-as" operators of the form $\overset{x}{\rhd}$. Similar properties of "counts-as" have been isolated in [Jones and Sergot, 1996] and [Grossi et al., 2006].

The invalidities 8a-8e show that operators $\overset{x}{\rhd}$ do not satisfy reflexivity (invalidity 8a), transitivity (invalidity 8b), strengthening of the antecedent (invalidity 8c), weakening of the consequent (invalidity 8d) and cautious monotonicity (invalidity 8e).

On the contrary, operators $\overset{x}{\rhd}$ satisfy the properties of right logical equivalence (Theorem 9a), left logical equivalence (Theorem 9b), conjunction of the consequents (Theorem 9c), disjunction of the antecedents (Theorem 9d), cumulative transitivity (Theorem 9e).

Theorem 8.

(8a) $\nvdash_{\mathcal{AL}} \varphi \overset{x}{\rhd} \varphi$

(8b) $\nvdash_{\mathcal{AL}} ((\varphi_1 \overset{x}{\rhd} \varphi_2) \wedge (\varphi_2 \overset{x}{\rhd} \varphi_3)) \rightarrow (\varphi_1 \overset{x}{\rhd} \varphi_3)$

(8c) $\nvdash_{\mathcal{AL}} (\varphi_1 \overset{x}{\rhd} \varphi_2) \rightarrow ((\varphi_1 \wedge \varphi_3) \overset{x}{\rhd} \varphi_2)$

(8d) $\nvdash_{\mathcal{AL}} (\varphi_1 \overset{x}{\rhd} \varphi_2) \rightarrow (\varphi_1 \overset{x}{\rhd} (\varphi_2 \vee \varphi_3))$

(8e) $\nvdash_{\mathcal{AL}} ((\varphi_1 \overset{x}{\rhd} \varphi_2) \wedge (\varphi_1 \overset{x}{\rhd} \varphi_3)) \rightarrow ((\varphi_1 \wedge \varphi_2) \overset{x}{\rhd} \varphi_3)$

Theorem 9. *For every $x \in INST$:*

(9a) *From* $\vdash_{\mathcal{AL}} (\varphi_2 \leftrightarrow \varphi_3)$ *infer* $\vdash_{\mathcal{AL}} (\varphi_1 \overset{x}{\rhd} \varphi_2) \leftrightarrow (\varphi_1 \overset{x}{\rhd} \varphi_3)$

(9b) *From* $\vdash_{\mathcal{AL}} (\varphi_1 \leftrightarrow \varphi_3)$ *infer* $\vdash_{\mathcal{AL}} (\varphi_1 \overset{x}{\rhd} \varphi_2) \leftrightarrow (\varphi_3 \overset{x}{\rhd} \varphi_2)$

(9c) $\vdash_{\mathcal{AL}} ((\varphi_1 \overset{x}{\rhd} \varphi_2) \wedge (\varphi_1 \overset{x}{\rhd} \varphi_3)) \rightarrow (\varphi_1 \overset{x}{\rhd} (\varphi_2 \wedge \varphi_3))$

(9d) $\vdash_{\mathcal{AL}} ((\varphi_1 \overset{x}{\rhd} \varphi_2) \wedge (\varphi_3 \overset{x}{\rhd} \varphi_2)) \rightarrow ((\varphi_1 \vee \varphi_3) \overset{x}{\rhd} \varphi_2)$

(9e) $\vdash_{\mathcal{AL}} ((\varphi_1 \overset{x}{\rhd} \varphi_2) \wedge ((\varphi_1 \wedge \varphi_2) \overset{x}{\rhd} \varphi_3)) \rightarrow (\varphi_1 \overset{x}{\rhd} \varphi_3)$

The invalidities 8a-8e are due to the local nature of the "counts-as" conditional $\varphi \overset{x}{\rhd} \psi$. For instance, the fact that $\varphi_1 \overset{x}{\rhd} \varphi_2$ and $\varphi_2 \overset{x}{\rhd} \varphi_3$ are constitutive rules of the institution x does not necessarily entail that $\varphi_1 \overset{x}{\rhd} \varphi_3$ is a constitutive rule of x since it does not necessarily entail $\neg [Univ] (\varphi_1 \rightarrow \varphi_3)$. This is the reason why $\overset{x}{\rhd}$ fails to satisfy transitivity.

Example 8. *In the US state of Texas, "to commit a murder counts as to be punishable by the Death Penalty", and "to be punishable by the Death Penalty counts as to be liable to indictment". As the Death Penalty is not universally accepted in all institutions, both these rules are constitutive rules of Texas,* i.e. $murder \overset{Texas}{\rhd} DeathPenalty$ *and* $DeathPenalty \overset{Texas}{\rhd} indictable$. *From this, it does not follow that it is a constitutive rule of Texas that "to commit a murder counts as to be liable to indictment". Indeed,* $\neg(murder \overset{Texas}{\rhd} indictable)$ *is true. This is due to the fact that "to commit a murder counts as to be liable to indictment" in all countries and institutions, and it is not constitutive of Texas,* i.e. $[Univ] (murder \rightarrow indictable)$.

Similarly, $\overset{x}{\rhd}$ fails to satisfy reflexivity. Indeed, all agents in all possible institutions accept the tautology $\varphi \rightarrow \varphi$ so that "φ counts as φ" cannot be intrinsically contextual

with respect to a certain institution. For similar reasons, strengthening of the antecedent is not a valid property of the operator $\overset{x}{\rhd}$.[3] The following example clarifies this aspect.

Example 9. *It is an accepted custom in the US that a person must leave a tip to the waiter that served him/her at a restaurant. That is, it is a constitutive rule of US that "not leaving a tip to the waiter counts as a violation",* i.e. $\neg leaveTip \overset{US}{\rhd} viol$. *From this, it does not follow that it is a constitutive rule of US that "not leaving a tip to the waiter and not paying the bill counts as a violation". Indeed,* $\neg((\neg leaveTip \wedge \neg payBill) \overset{US}{\rhd} viol)$ *is true. This is because "not leaving a tip and not paying the bill counts as a violation" in all countries and institutions, and it is not constitutive of US,* i.e. $[Univ]((\neg leaveTip \wedge \neg payBill) \rightarrow viol)$.

Discussion

The formal analysis of "counts-as" presented in this section is in agreement with the formal analysis of "counts-as" proposed in [Grossi et al., 2006], where a notion of *proper classificatory rule* is introduced. A proper classificatory rule is represented by the construction $\varphi \Rightarrow_x^{cl+} \psi$ which is meant to stand for "φ counts as ψ in the normative system x". Proper classificatory rules are distinguished by Grossi et al. from (non-proper) *classificatory rules* of type $\varphi \Rightarrow_x^{cl} \psi$. In a way similar to our concept of constitutive rule, *proper classificatory rules* have the specific property of not being universally valid (*i.e.* valid in all institutional contexts). That is, differently from non-proper *classificatory rules*, *proper classificatory rules* are rules which would not hold without the normative system/institution stating them.[4]

Non-proper classificatory rules could be expressed in our logical framework by constructions of the form $[x](\varphi \rightarrow \psi)$, that is, by removing the condition $\neg[Univ](\varphi \rightarrow \psi)$ from the definition of $\varphi \overset{x}{\rhd} \psi$. In agreement with Grossi et al., we would be able to prove that, differently from constitutive rules of the form $\varphi \overset{x}{\rhd} \psi$, such a kind of rules satisfy reflexivity, transitivity, strengthening of the antecedent, weakening of the consequent, and cautious monotonicity. Indeed, the following formulas are all theorems of our logic $\mathcal{AL}$:

Theorem 10.

(10a) $\vdash_{\mathcal{AL}} [x](\varphi \rightarrow \varphi)$

(10b) $\vdash_{\mathcal{AL}} ([x](\varphi_1 \rightarrow \varphi_2) \wedge [x](\varphi_2 \rightarrow \varphi_3)) \rightarrow [x](\varphi_1 \rightarrow \varphi_3)$

(10c) $\vdash_{\mathcal{AL}} [x](\varphi_1 \rightarrow \varphi_2) \rightarrow [x]((\varphi_1 \wedge \varphi_3) \rightarrow \varphi_2)$

(10d) $\vdash_{\mathcal{AL}} [x](\varphi_1 \rightarrow \varphi_2) \rightarrow [x](\varphi_1 \rightarrow (\varphi_2 \vee \varphi_3))$

(10e) $\vdash_{\mathcal{AL}} ([x](\varphi_1 \rightarrow \varphi_2) \wedge [x](\varphi_1 \rightarrow \varphi_3)) \rightarrow [x]((\varphi_1 \wedge \varphi_2) \rightarrow \varphi_3)$

In Section 8.1 a more elaborate and detailed analysis of the logic presented in [Grossi et al., 2006] will be provided and its formal relationships with our logic of acceptance will be studied.

[3] Other authors have defended the idea that strengthening of the antecedent and transitivity should not be valid properties of "counts-as" conditionals (*e.g.* [Gelati et al., 2004]).

[4] See [Grossi et al., 2008] for a refinement of this typology of rules.

Another important aspect to be discussed about our formalization of "counts-as" is the problem of contraposition. Indeed, at the present stage, $\varphi \overset{x}{\rhd} \psi$ is logically equivalent to $\neg\psi \overset{x}{\rhd} \neg\varphi$ which can be counterintuitive in some situations. However, the problem of contraposition could be solved by distinguishing in the language of the logic $\mathcal{AL}$ formulas denoting "brute" physical facts from formulas denoting institutional facts and by imposing that the consequent ψ of a "counts-as" conditional $\varphi \overset{x}{\rhd} \psi$ is always a formula denoting an institutional fact. Under this assumption, if the negation of the antecedent in the "counts-as" conditional is not an institutional fact (*i.e.* formula $\neg\varphi$ does not denote an institutional fact), contraposition is not allowed. That is, $\varphi \overset{x}{\rhd} \psi$ does not imply $\neg\psi \overset{x}{\rhd} \neg\varphi$.[5] It is worth noting that, this distinction between formulas denoting "brute" physical facts and formulas denoting institutional facts would enable us to account for an aspect of "counts-as" that our current formalization is not able to capture, namely: the function of "counts-as" statements of establishing the relations between physical facts and objects on the one hand (the antecedent of the "counts-as"), and institutional facts and objects on the other hand (the consequent of the "counts-as"), *e.g.* a certain piece of paper counts as a five-euro bill.

6.2 Regulative rules

Constitutive rules as defined in the previous Section 6.1 are still not sufficient for a characterization of institutional reality. An institution is indeed connected to a deontic dimension that up to now is still missing in our analysis. This deontic dimension consists in several concepts such as obligation, permission, prohibition, *etc.* which are aimed at regulating agents' behaviors and social interactions within the context of the institution.

In order to capture this deontic dimension of institutions, our logic $\mathcal{AL}$ can be appropriately extended by introducing a *violation* atom $viol$ as in Anderson's reduction of deontic logic to alethic logic [Anderson, 1958] and in dynamic deontic logic [Meyer, 1988]. A similar approach has been recently taken in [Grossi, 2008]. By means of the new formal construct $viol$ we can specify the concepts of obligation and that of permission in a way that respects their being also a kind of attitude-dependent facts holding in a specific institutional context.

As far as obligations are concerned, we introduce operators of the form O_x which are used to specify what is obligatory in the context of a certain institution x:

$$O_x\varphi \overset{def}{=} \neg\varphi \overset{x}{\rhd} viol$$

According to this definition, "φ is obligatory within the institutional context x" if and only if "$\neg\varphi$ counts as a violation within the institutional context x".

Example 10. *The formula* $(driveCar \wedge RightSide) \overset{UK}{\rhd} viol$ *which is equivalent to* $O_{UK}(driveCar \rightarrow \neg RightSide)$ *expresses that in the UK it is obligatory to drive on*

[5] See also [Grossi, 2008] for a different solution on how to solve the problem of contraposition in a normal modal logic of "counts-as".

*the left side of the street (*i.e. *"driving a car on the right side of the street counts as violation in UK").*

As the following theorem highlights, our O_x operators satisfy axiom K (Theorem 11a) and do not allow obligations about tautologies (Theorem 11b).

Theorem 11. *For every $x \in INST$:*

(11a) $\vdash_{\mathcal{AL}} \quad O_x(\varphi \rightarrow \psi) \rightarrow (O_x\varphi \rightarrow O_x\psi)$

(11b) $\vdash_{\mathcal{AL}} \neg O_x \top$

On the contrary, obligation operators do not satisfy the necessitation rule. This is due to the negative condition $\neg\,[\,Univ\,]\,(\neg\varphi \rightarrow viol)$ in the definition of $O_x\varphi$. Indeed, in order to have a normal modal operator for obligation, it is sufficient to remove the negative condition $\neg\,[\,Univ\,]\,(\varphi \rightarrow \psi)$ from the definition of the "counts-as" conditional $\varphi \overset{x}{\triangleright} \psi$ given in Section 6.1. The following theorem highlights other interesting invalidities of the obligation operators O_x.

Theorem 12.

(12a) $\nvdash_{\mathcal{AL}} \neg O_x \bot$

(12b) $\nvdash_{\mathcal{AL}} O_x\varphi \rightarrow O_x(\varphi \vee \psi)$

(12c) $\nvdash_{\mathcal{AL}} O_x(\varphi \wedge \psi) \rightarrow O_x\varphi$

According to the invalidity 12a, obligation operators do not satisfy the axiom D of Standard Deontic Logic (SDL) [Åqvist, 2002]. For instance, in the logic $\mathcal{AL}$ institutions might be empty, that is, for every $C \in 2^{AGT^\star}$, $\mathcal{A}_{C:x}\bot$. If institution x is empty, it does not have any obligation (*i.e.* $O_x\bot$). According to the other two invalidities we have that: if φ is obligatory within the context of institution x then, it is not necessarily the case that φ or ψ is obligatory within the context of the same institution (invalidity 12b) and if φ and ψ are obligatory within the context of institution x then, it is not necessarily the case that φ is obligatory within the context of the same institution (invalidity 12c). Thus, our obligation operators O_x do not incur two classical problems of Standard Deontic Logic which are commonly referred to as "Ross paradox" and "Good Samaritan paradox" [Carmo and Jones, 2002]. On the one hand, it seems rather odd to say that the *obligation to mail a certain letter* entails an *obligation to mail the letter or to burn it* which can be fulfilled simply by burning the letter (something presumably forbidden) ("Ross paradox"). On the other hand, it seems rather odd to say that if *it is obligatory that Mary helps John who has had an accident*, then *it is obligatory that John has an accident* ("Good Samaritan paradox"). Here we do not consider other well-known paradoxes of deontic logic (such as Chisholm paradox for instance) which require an elaborate and detailed analysis of contrary-to-duty obligations and defeasible conditional obligations (on this see [Prakken and Sergot, 1997, Hansen et al., 2007] for instance). Indeed, this issue goes beyond the objectives of the present work.

As far as permissions are concerned we say that "φ is permitted within the institutional context x" (noted $P_x\varphi$) if and only if $\neg\varphi$ is not obligatory within the institutional context x. Formally:

$$P_x\varphi \overset{def}{=} \neg O_x \neg\varphi$$

That is, we define the permission operator in the standard way as the dual of the obligation operator.[6]

Before concluding this section, it is important to stress again that in our approach regulative rules of type $O_x\varphi$ and $P_x\varphi$ as well as constitutive rules of type $\varphi \overset{x}{\triangleright} \psi$ of a certain institution are attitude-dependent facts which are grounded on the acceptances of the members of a certain institution.

7 Towards legal institutions

In Section 5 we have supposed that φ is true within the context of institution x if and only if all members of this institution accept φ to be true. At this point, it might be objected that there are facts which are true in an institutional context but only "special" members of the institution are aware of them. For instance, there are laws in every country that are known only by the specialists of the domain (lawyers, judges, members of the Parliament, *etc.*). Aren't these facts true notwithstanding that many members of the institution are not aware of them?

In order to resist to this objection recall that until now our model applied to the basic informal institutions of a society, that is, *rule-governed social practices* [Tuomela, 2002] in which no member with "special" powers is introduced.

It is a peculiar property of informal institutions to be based on general consensus [Coleman, 1990], that is, a certain fact φ is true within the context of an informal institution x if and only if all members of x accept φ to be true. Relative to this restriction, the assumption made in Section 5 is justified because, with respect to informal institutions, there are no specialized agents called legislators empowered to change the institution itself on behalf of everybody else. For instance, in the informal institution of common language, nobody has the power to change the rules for promising. (See [Searle, 1969] for more details.) On the contrary, it is a specificity of legal (formal) institutions to have such specialized agents with special powers to interpret and modify the institution itself. This distinction between informal and formal (legal) institutions has been stressed by many authors working in the field of social and legal theory [Castelfranchi, 2003, North, 1990, Lorini and Longin, 2008, Von Wright, 1963]. Consider for instance the following quotation from Von Wright where the terms *prescription* and *custom* respectively correspond to the terms *formal institution* and *informal institution* used here: "(...) Prescriptions are *given* or *issued* by someone. They 'flow' from or have their 'source' in the will of norm-giver (...) Customs, first of all, are not *given* by any authority to subjects. If we can speak of an authority behind the customs at all this authority would be the community itself" [Von Wright, 1963, p. 7–9].

In the rest of this section we will show how the logic $\mathcal{AL}$ can be appropriately refined in order to move beyond informal institutions and to capture some essential

[6] We do not consider here the classical distinction between *weak permission* and *strong permission* [Alchourrón and Bulygin, 1971, Raz, 1975, Von Wright, 1963]. According to legal theory, a weak permission corresponds to the absence in a normative system of a norm prohibiting φ (this is represented by our permission operator P_x). A strong permission corresponds to the existence in the normative system of an explicit norm, issued by the legislators, according to which φ is permitted. For a logical analysis of the distinction between weak and strong permission see our related work [Lorini and Longin, 2008].

properties of formal (legal) institutions in which legislators are introduced. We will discuss some general principles which seem adequate for a formal characterization of legal institutions. For the sake of simplicity and readability of the article, these principles will not be included in the axiomatization of the logic $\mathcal{AL}$ and their semantic counterparts will not be studied.

In order to distinguish formal from informal institutions, we introduce a total function Leg which assigns a (possibly empty) set of agents to every institution x:

$$Leg : INST \longrightarrow 2^{AGT}$$

$Leg(x)$ denotes the set of legislators of institution x, that is, the set of agents legally responsible over institution x and which are entitled to modify its structure. The function Leg allows distinguishing formal from informal institutions in a simple way. It is indeed reasonable to suppose that informal institutions are those institutions that do not have legislators, that is, x is an informal institution if and only if $Leg(x) = \emptyset$. On the contrary, if $Leg(x) \neq \emptyset$, x is a legal or formal institution. In this sense, the cardinality of $Leg(x)$ provides an important property: it allows us to distinguish between legal institutions and informal institutions.

It seems reasonable to suppose that the legislators of a certain legal institution x must function together as members of institution x. This assumption is expressed by the following principle. For any $x \in INST$ such that $Leg(x) \neq \emptyset$:

$$\neg \mathcal{A}_{Leg(x):x} \bot$$

As emphasized in Section 5, legislators are "special" agents who have the power to affect the acceptances of the other members of the institution. In legal institutions, all facts that are accepted by the legislators must be universally accepted by all members of the institution. In this perspective, legal institutions are characterized by the following principle which explains how the collective acceptance of a set C of members of institution x is affected by the acceptance of the legislators of the institution. For every $C \in 2^{AGT\star}$ and $x \in INST$ such that $Leg(x) \neq \emptyset$:

(**Legislators**) $$\mathcal{A}_{C:x}(\bigwedge_{i \in Leg(x)} \mathcal{A}_{i:x}\varphi \rightarrow \varphi)$$

According to **Legislators**, for every group of agents C, while functioning together as members of the institution x, the agents in C accept that if the legislators of x accept that φ, then φ is the case. As emphasized in Section 4.1, the Principle **Legislators** can be conceived as an additional specification of how collective acceptances of groups of agents are built within the context of an institution. It is worth noting that **Legislators** is perfectly compatible with the general principle of unanimity of the logic $\mathcal{AL}$ described by Axiom **Unanim** (and the related Theorems 3e, 3f). Indeed, we can reasonably suppose that the members of an institution might accept certain things on the basis of a criterion of unanimity and, at the same time, accept what the legislators accept and decide.[7]

[7]Note that a further principle which seems reasonable for legal institutions is a majority principle for

We conclude by showing how the concept of institutional truth proposed in Section 5 can be appropriately refined in order to deal with legal institutions. Differently from informal institutions, legal institutions do not necessarily depend on the general consensus of all their members. More precisely, if a certain fact φ is true within the context of the legal institution x then, it is not necessarily the case that for every set of agents C, the agents in C accept φ while functioning together as members of the legal institution x. In a legal institution it is sufficient that the legislators accept φ to be true to make it true for the institution. This means that the notion of institutional truth for legal institutions should be defined as follows. For any $x \in INST$ such that $Leg(x) \neq \emptyset$:

$$[x]^L \varphi \stackrel{def}{=} \mathcal{A}_{Leg(x):x}\varphi$$

This means that "within the context of the legal institution x it is the case that φ" if and only if "the legislators of institution x accept that φ".

From the principles of $\mathcal{AL}$ and the definition of the function $Leg()$, it follows that the operators $[x]^L$ are also normal. Moreover, differently from the $[x]$ operators, which adequately characterize the notion of institutional truth for informal institutions, $[x]^L$ operators satisfy axioms 4 and 5 of modal logic, that is: if the legislators of institution x accept φ then, they accept that they accept φ (Theorem 13c); if the legislators of an institution x do not accept φ, then they accept that they do not accept φ (Theorem 13d).[8]

Theorem 13. *For every $x \in INST$:*

(13a) $\quad \vdash_{\mathcal{AL}} \quad [x]^L (\varphi \to \psi) \to ([x]^L \varphi \to [x]^L \psi)$

(13b) $\quad$ *From* $\vdash_{\mathcal{AL}} \varphi$ *infer* $\vdash_{\mathcal{AL}} [x]^L \varphi$

(13c) $\quad \vdash_{\mathcal{AL}} \quad [x]^L \varphi \to [x]^L [x]^L \varphi$

(13d) $\quad \vdash_{\mathcal{AL}} \quad \neg [x]^L \varphi \to [x]^L \neg [x]^L \varphi$

It is worth noting that the analysis of constitutive rules and regulative rules proposed in Sections 6.1 and 6.2 could be refined in the light of this distinction between informal and legal institutions. In particular, a new form of "counts-as" and two related concepts of obligation and permission could be defined in terms of the previous operator $[x]^L$. This is in order to characterize a notion of constitutive rule and a notion of regulative rule which apply straightforwardly to the context of legal institutions, and which go beyond the notions of constitutive rule and regulative rule for informal institutions given in Sections 6.1 and 6.2 and based on the operator $[x]$. We postpone this kind of analysis to future works.

---

legislators: the legislators of a certain legal institution x accept that if the majority of them accept φ, then φ is true. This should be conceived as a particular case of the majority principle discussed in Section 4.1. Formally, for any $x \in INST$ such that $Leg(x) \neq \emptyset$, if $B \subseteq Leg(x)$ and $|Leg(x) \setminus B| < |B|$ (*i.e.* B represents the majority of the legislators of the institution x) then: $\mathcal{A}_{Leg(x):x}(\bigwedge_{i \in B} \mathcal{A}_{i:x}\varphi \to \varphi)$.

[8] Note that the operator $[x]$ is stronger than the operator $[x]^L$, that is, $[x]\,\varphi$ implies $[x]^L\,\varphi$.

8 Comparison with other logical approaches to normative systems

In the following two sections our logic $\mathcal{AL}$ will be compared with two approaches to normative systems and institutions which have been recently proposed in the multi-agent system domain.

8.1 Embedding Grossi et al.'s logic of "counts-as" into $\mathcal{AL}$

Because of the interesting formal similarities, we will first compare $\mathcal{AL}$ with the modal logic of normative systems proposed in [Grossi et al., 2006], henceforth abbreviated $\mathcal{GMD}$ logic.

In the $\mathcal{GMD}$ logic a set of contexts CXT denoting normative systems is introduced. $\mathcal{GMD}$ logic is based on a set of modal operators $[\![x]\!]$ (one for every context x in CXT). Operators $[\![x]\!]$ are similar to our operators $[x]$ defined in Section 5.[9] A formula $[\![x]\!]\,\varphi$ approximately stands for "in the institutional context/normative system x it is the case that φ". It is supposed that CXT contains a special context $Univ$, where the operator $[\![Univ]\!]$ is used for denoting facts which universally hold. We note $CXT_0 = CXT \setminus \{Univ\}$. The language of the $\mathcal{GMD}$ logic is given by the following BNF:

$$\varphi \ ::= \ \bot \mid p \mid \neg\varphi \mid \varphi \vee \varphi \mid [\![x]\!]\,\varphi$$

where p ranges over ATM and x ranges over CXT. $\wedge$, $\rightarrow$, $\leftrightarrow$ and $\top$ are defined from $\vee$, $\neg$ and $\bot$ in the usual manner.

As noted in Section 6.1, operators $[\![x]\!]$ and $[\![Univ]\!]$ are exploited in Grossi et al.'s logic to define contextual conditionals called *proper classificatory rules*, noted $\varphi \Rightarrow_x^{cl+} \psi$, which are an abbreviation of $[\![x]\!]\,(\varphi \rightarrow \psi) \wedge \neg\,[\![Univ]\!]\,(\varphi \rightarrow \psi)$ and which read "φ counts as ψ in the normative system x". The construction $\varphi \Rightarrow_x^{cl+} \psi$ is similar to our $\varphi \overset{x}{\rhd} \psi$.

The most striking difference between our logic of acceptance $\mathcal{AL}$ and the $\mathcal{GMD}$ logic is that in the logic $\mathcal{AL}$ the contextual operators $[x]$ are built on the notion of collective acceptance, whereas in the $\mathcal{GMD}$ logic the contextual operators $[\![x]\!]$ are given as primitive operators.

Frames of the $\mathcal{GMD}$ logic are called *multi-context frames*. A multi-context frame has the following form:

$$\mathcal{F}^{\mathcal{GMD}} = \langle S, \{S_x\}_{x \in CXT_0} \rangle$$

where:

- S is a set of possible worlds;
- $\{S_x\}_{x \in CXT_0}$ is a family of subsets of S, one for every institutional context $x \in CXT_0$.

A *multi-context model* is a tuple

$$\mathcal{M}^{\mathcal{GMD}} = \langle \mathcal{F}^{\mathcal{GMD}}, \pi \rangle$$

where:

[9] Here we use the notation $[\![x]\!]$ in order to distinguish their operators from ours.

- $\mathcal{F}^{\mathcal{GMD}}$ is a multi-context frame;
- $\pi : ATM \rightarrow 2^S$ is a valuation function associating a set of possible worlds $\pi(p) \subseteq S$ to each atomic formula p of ATM.

The truth conditions for formulas of the $\mathcal{GMD}$ logic are just standard for contradiction, atomic propositions, negation and disjunction. The following are the truth conditions for $[\![x]\!]\,\varphi$ and $[\![Univ]\!]\,\varphi$.

- $\mathcal{M}^{\mathcal{GMD}}, w \models [\![x]\!]\,\varphi$ iff $\mathcal{M}, w' \models \varphi$ for all $w' \in S_x$;
- $\mathcal{M}^{\mathcal{GMD}}, w \models [\![Univ]\!]\,\varphi$ iff $\mathcal{M}, w' \models \varphi$ for all $w' \in S$.

A formula φ is *true in a $\mathcal{GMD}$ model* $\mathcal{M}^{\mathcal{GMD}}$ iff $\mathcal{M}^{\mathcal{GMD}}, w \models \varphi$ for every world w in $\mathcal{M}^{\mathcal{GMD}}$. φ is $\mathcal{GMD}$ *valid* (noted $\models_{\mathcal{GMD}} \varphi$) if and only if φ is true in all $\mathcal{GMD}$ models. φ is $\mathcal{GMD}$ *satisfiable* iff $\neg\varphi$ is not $\mathcal{GMD}$ valid.

The $\mathcal{GMD}$ logic is axiomatized by the following principles, where x and y denote elements of the set CXT_0 :

(**ProTau**)	All principles of propositional calculus
($\mathbf{K}_{[\![x]\!]}$)	$[\![x]\!]\,(\varphi \rightarrow \psi) \rightarrow ([\![x]\!]\,\varphi \rightarrow [\![x]\!]\,\psi)$
($\mathbf{K}_{[\![Univ]\!]}$)	$[\![Univ]\!]\,(\varphi \rightarrow \psi) \rightarrow ([\![Univ]\!]\,\varphi \rightarrow [\![Univ]\!]\,\psi)$
($\mathbf{4}_{[\![x]\!],[\![y]\!]}$)	$[\![x]\!]\,\varphi \rightarrow [\![y]\!]\,[\![x]\!]\,\varphi$
($\mathbf{5}_{[\![x]\!],[\![y]\!]}$)	$\neg\,[\![x]\!]\,\varphi \rightarrow [\![y]\!]\,\neg\,[\![x]\!]\,\varphi$
($\mathbf{4}_{[\![Univ]\!]}$)	$[\![Univ]\!]\,\varphi \rightarrow [\![Univ]\!]\,[\![Univ]\!]\,\varphi$
($\mathbf{5}_{[\![Univ]\!]}$)	$\neg\,[\![Univ]\!]\,\varphi \rightarrow [\![Univ]\!]\,\neg\,[\![Univ]\!]\,\varphi$
($\mathbf{T}_{[\![Univ]\!]}$)	$[\![Univ]\!]\,\varphi \rightarrow \varphi$
($\subseteq_{[\![Univ]\!],[\![x]\!]}$)	$[\![Univ]\!]\,\varphi \rightarrow [\![x]\!]\,\varphi$
(**MP**)	From $\vdash \varphi$ and $\vdash \varphi \rightarrow \psi$ infer $\vdash \psi$
($\mathbf{Nec}_{[\![x]\!]}$)	From $\vdash \varphi$ infer $\vdash [\![x]\!]\,\psi$
($\mathbf{Nec}_{[\![Univ]\!]}$)	From $\vdash \varphi$ infer $\vdash [\![Univ]\!]\,\psi$

We write $\vdash_{\mathcal{GMD}} \varphi$ if formula φ is a theorem of $\mathcal{GMD}$.

Axiom $\mathbf{K}_{[\![x]\!]}$ and Rule $\mathbf{Nec}_{[\![x]\!]}$ express that the operators $[\![x]\!]$ are normal modal operators. Axioms $\mathbf{K}_{[\![Univ]\!]}$, $\mathbf{4}_{[\![Univ]\!]}$, $\mathbf{5}_{[\![Univ]\!]}$, $\mathbf{T}_{[\![Univ]\!]}$ and the rule of inference $\mathbf{Nec}_{[\![Univ]\!]}$ express that the universal modality $[\![Univ]\!]$ is defined in the modal logic system S5. According to the Axioms $\mathbf{4}_{[\![x]\!],[\![y]\!]}$ and $\mathbf{5}_{[\![x]\!],[\![y]\!]}$, truth and falsehood in institutional contexts/normative systems are absolute because they remain invariant even if they are evaluated from another institutional context/normative system. This means that every normative system y has full access to all facts which are true in a different normative system x. In our view, these two principles are criticizable because they rely on a strong assumption of perfect information, *i.e.* a normative system has perfect information about the facts that are true in the other normative systems. Axiom $\subseteq_{[\![Univ]\!],[\![x]\!]}$ expresses the relationship between the universal modality and the contextual modalities.

In [Grossi et al., 2006] it is proved that the $\mathcal{GMD}$ logic is sound and complete with respect to the class of $\mathcal{GMD}$ frames.

It is easy to show that the principles of the acceptance logic $\mathcal{AL}$ given in Section 3 are not sufficient to derive the principles of the $\mathcal{GMD}$ logic. In particular, Axioms $\mathbf{4}_{[\![x]\!],[\![y]\!]}$, $\mathbf{5}_{[\![x]\!],[\![y]\!]}$, $\mathbf{4}_{[\![Univ]\!]}$, $\mathbf{5}_{[\![Univ]\!]}$ and $\mathbf{T}_{[\![Univ]\!]}$ are not derivable in $\mathcal{AL}$.

In order to embed $\mathcal{GMD}$ we need to slightly modify the properties of the logic $\mathcal{AL}$. On the one hand, we need to generalize Axioms **PAccess** and **NAccess** by supposing that they **also** hold for the case $B \not\subseteq C$. This is in order to infer the formulas $[x]\,\varphi \rightarrow [y]\,[x]\,\varphi$ and $\neg\,[x]\,\varphi \rightarrow [y]\,\neg\,[x]\,\varphi$ in the augmented logic $\mathcal{AL}$. Thus, we need to assume that, given two arbitrary sets of agents B and C, the agents in B have access to all facts that the agents in C accept (do not accept), while functioning together as members of a certain institution x. On the other hand, we need to add the principle $[Univ]\,\varphi \rightarrow \varphi$ to the logic $\mathcal{AL}$. The way to embed the $\mathcal{GMD}$ logic into our logic $\mathcal{AL}$ is illustrated in the following paragraph.

An embedding of $\mathcal{GMD}$ logic. Let us slightly modify the logic of acceptance $\mathcal{AL}$ in order to provide a correct embedding of $\mathcal{GMD}$. We call $\mathcal{AL}^+$ the modified logic of acceptance.

$\mathcal{AL}^+$ has the same language as $\mathcal{AL}$ (see Section 3.1). $\mathcal{AL}^+$ frames are tuples $\mathcal{F} = \langle W, \mathscr{A} \rangle$ where W and $\mathscr{A}$ are defined as for $\mathcal{AL}$ frames, except that the constraints **S.1** and **S.2** given in Section 3.2 are supposed to hold also for the case $B \not\subseteq C$ and the following additional constraint **S.6** is imposed. That is, for any world $w \in W$, institutional context $x \in INST$, and sets of agents $C, B \in 2^{AGT\star}$ we suppose:

(S.1') if $w' \in \mathscr{A}_{B:y}(w)$ then $\mathscr{A}_{C:x}(w') \subseteq \mathscr{A}_{C:x}(w)$

(S.2') if $w' \in \mathscr{A}_{B:y}(w)$ then $\mathscr{A}_{C:x}(w) \subseteq \mathscr{A}_{C:x}(w')$

Furthermore, for any world $w \in W$ we suppose:

(S.6) $\exists C \in 2^{AGT\star}$, $\exists x \in INST$ such that $w \in \mathscr{A}_{C:x}(w)$

The axiomatization of $\mathcal{AL}^+$ is given by the axiom schemes and rules of inference of $\mathcal{AL}$, except that an Axiom corresponding to the Axiom $\mathbf{T}_{[\![Univ]\!]}$ of the $\mathcal{GMD}$ logic is added, and the Axioms **PAccess** and **NAccess** of the logic $\mathcal{AL}$ are generalized in such a way that they also for hold for the case $B \not\subseteq C$. That is, for any sets of agents $C, B \in 2^{AGT\star}$, we suppose:

(PAccess$^+$) $\mathcal{A}_{C:x}\varphi \rightarrow \mathcal{A}_{B:y}\mathcal{A}_{C:x}\varphi$

(NAccess$^+$) $\mathcal{A}_{C:x}\varphi \rightarrow \mathcal{A}_{B:y}\neg\mathcal{A}_{C:x}\varphi$

Furthermore, we suppose:

($\mathbf{T}_{Univ}$) $[Univ]\,\varphi \rightarrow \varphi$

Axioms **PAccess**$^+$ and **NAccess**$^+$ respectively correspond to the semantic constraints **S.1'** and **S.2'**, whilst Axiom $\mathbf{T}_{Univ}$ corresponds to the semantic constraint **S.6**.

The definitions of validity and satisfiability in $\mathcal{AL}^+$ are given accordingly. We write $\models_{\mathcal{AL}^+} \varphi$ if formula φ is *valid* in all $\mathcal{AL}^+$ models satisfying the semantic constraints **S.3**, **S.4**, **S.5** given in Section 3.2 and the constraints **S.1'**, **S.2'**, **S.6** given here. We call $\mathcal{AL}^+$ the logic axiomatized by Axiom $\mathbf{T}_{Univ}$ and the principles of the logic $\mathcal{AL}$ (Section 3.4), where Axioms **PAccess** and **NAccess** are generalized to **PAccess**$^+$ and **NAccess**$^+$. We write $\vdash_{\mathcal{AL}^+} \varphi$ if formula φ is a theorem of $\mathcal{AL}^+$.

We can prove that $\mathcal{AL}^+$ as well is sound and complete. More precisely:

Theorem 14. *$\vdash_{\mathcal{AL}^+} \varphi$ if and only if $\models_{\mathcal{AL}^+} \varphi$.*

Consider the following translation tr from $\mathcal{GMD}$ to the new logic $\mathcal{AL}^+$:

- $tr(\bot) = \bot$
- $tr(p) = p$
- $tr(\neg\varphi) = \neg tr(\varphi)$
- $tr(\varphi \vee \psi) = tr(\varphi) \vee tr(\psi)$
- $tr([\![x]\!]\,\varphi) = [x]\,tr(\varphi)$
- $tr([\![Univ]\!]\,\varphi) = [Univ]\,tr(\varphi)$,

As the following Theorem 15 shows, tr is a correct embedding of the $\mathcal{GMD}$ logic.

Theorem 15. *Let $INST = CXT$ and φ be a formula of the $\mathcal{GMD}$ logic. Then, φ is $\mathcal{GMD}$ satisfiable if and only if $tr(\varphi)$ is $\mathcal{AL}^+$ satisfiable.*

REMARK. It is worth noting that $\mathcal{GMD}$ logic can also be embedded into the variant of $\mathcal{AL}$ with legislators presented in Section 7 by the translations $tr([\![x]\!]\,\varphi) = [x]^L\,\varphi$ and $tr([\![Univ]\!]\,\varphi) = [Univ]^L\,\varphi$, after defining

$$[Univ]^L\,\varphi \stackrel{def}{=} \bigwedge_{x \in INST} [x]^L\,\varphi.$$

To obtain a correct embedding of the $\mathcal{GMD}$ logic, it is sufficient to add to $\mathcal{AL}$ the three axioms $[x]^L\,\varphi \rightarrow [y]^L\,[x]^L\,\varphi$, $\neg\,[x]^L\,\varphi \rightarrow [y]^L\,\neg\,[x]^L\,\varphi$ and $[Univ]^L\,\varphi \rightarrow \varphi$ and the two corresponding semantic constraints over $\mathcal{AL}$ frames:

$$\text{if } w' \in \mathscr{A}_{Leg(y):y}(w) \text{ then } \mathscr{A}_{Leg(x):x}(w') = \mathscr{A}_{Leg(x):x}(w), \text{ and}$$
$$\exists x \in INST \text{ such that } w \in \mathscr{A}_{Leg(x):x}(w).$$

8.2 A conceptual comparison with Boella & van der Torre's model

The formal approach to institutions and normative systems proposed by Boella & van der Torre [Boella and van der Torre 2004, 2007] is similar in some respect to ours. Here we just provide a *conceptual* comparison between the two approaches. We are not able to provide a more *technical* comparison. Indeed, our formalism based on modal logic and their formalism based on input-output logic [Makinson and van der Torre, 2000] are too different to be compared in the fashion followed in Section 8.1.

Boella & van der Torre emphasize the relevance of the concept of acceptance for a formal model of institutions. In their model, individual agents *accept* a norm, together

with its associated sanctions and rewards, when they recognize that this norm serves to achieve their desires and believe that the other agents will conform to it. According to them, for a norm to be really effective it must be respected due to its acceptance, and not only to the fear of sanctions. Although they take the concept of acceptance into consideration, they do not analyze it in detail. In particular, in their model there is no distinction between individual acceptance and collective acceptance. On the contrary, this distinction is fundamental in our $\mathcal{AL}$ logic in which we clarify the relationships between individual acceptances and collective acceptances and we provide an explanation of how the collective acceptance of a group of agents C is built from the individual acceptances of the agents in C.

Moreover, in Boella & van der Torre's approach, normative systems and institutions are conceived as agents and mental attitudes such as beliefs and goals are ascribed to them. Differently from them, we do not claim that institutions can be conceived as agents. In our approach, we only defend the idea that the institutional reality is built on the top of the agents' attitudes. In particular, we claim that institutions are grounded on the individual and collective *acceptances* of their members and groups of members, and their dynamics depend on the dynamics of these acceptances.

9 Conclusion

We have presented in this article a logic of acceptance and applied it to the analysis of institutions. Our logic of acceptance allows to express that agents accept something to be true *qua* members of a certain institution. Given the properties of this demystified notion of acceptance, we have provided an analysis of the kind of attitude-dependent facts which are typical of institutions. We have formalized the concept of constitutive rule expressed by statements of the form "X counts as Y in the context of institution x". Then, we have introduced a notion of obligation and a notion permission with respect to an institutional context (*i.e.* so-called regulative rules). While constitutive rules and regulative rules are usually defined from the external perspective of a normative system or institution, in the present work we have anchored these rules in the agents' acceptances.

Directions for future research are manifold. For instance, future works will be devoted to integrate modalities expressing agents' goals and preferences, such as the ones provided in [Cohen and Levesque, 1990], into the logical framework presented in this paper. This is in order to investigate the decision to join (resp. not to join) a given institution and the related decision to accept (resp. not to accept) the norms of the institution with its associated sanctions and rewards. These kinds of decisions are indeed influenced by the inconsistency between the agent's goals and the current norms and rules of the institution. For instance, if the agent's goals conflict with the norms proclaimed by the legislators then, the agent will probably decide not to join the institution.

Another interesting topic to be investigated in future works is the dynamics of individual and collective acceptances in institutional contexts. We have already started to study this topic in a recent work [Herzig et al., 2008]. The idea is to extend the logic of acceptance $\mathcal{AL}$ by events of type $x!\varphi$ and corresponding dynamic operators

of the form $[x!\varphi]$. A formula $[x!\varphi]\psi$, means that ψ is true after every announcement of formula φ in the context of institution x. Operators of type $[x!\varphi]$, which are similar to the operators of announcements in dynamic epistemic logic [Baltag et al., 1998, Gerbrandy and Groeneveld, 1997, van Ditmarsch et al., 2007], express that the members of an institution x learn that φ is true in that institution in such a way that their acceptances, *qua* members of institution x, are updated. Such operators can also be used to describe how the acceptances of the members of institution x change, after that a certain norm (e.g. obligation, permission) is *issued* or *promulgated* within the context of this institution.

References

[Ågotnes et al., 2007] Ågotnes, T., van der Hoek, W., Rodriguez-Aguilar, J., Sierra, C., and Wooldridge, M. (2007). On the logic of normative systems. In *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI'07)*, pages 1181–1186. AAAI Press.

[Alchourrón and Bulygin, 1971] Alchourrón, C. and Bulygin, E. (1971). *Normative systems*. Springer, New York.

[Anderson, 1958] Anderson, A. (1958). A reduction of deontic logic to alethic modal logic. *Mind*, 22:100–103.

[Åqvist, 2002] Åqvist, L. (2002). Deontic Logic. In Gabbay, D. and Guenthner, F., editors, *Handbook of Philosophical Logic*, volume 8, pages 147–264. Kluwer Academic Publishers, 2nd edition.

[Baltag et al., 1998] Baltag, A., Moss, L., and Solecki, S. (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)*, pages 43–56, San Francisco, CA. Morgan Kaufmann Publishers Inc.

[van Benthem, 2001] van Benthem, J. (2001). Correspondence theory. In Gabbay, D. and Guenthner, F., editors, *Handbook of Philosophical Logic*, volume 3, pages 325–408. Kluwer Academic Publishers, 2nd edition.

[Blackburn et al., 2001] Blackburn, P., de Rijke, M., and Venema, Y. (2001). *Modal Logic*. Cambridge University Press, Cambridge.

[Boella and van der Torre, 2004] Boella, G. and van der Torre, L. (2004b). Regulative and constitutive norms in normative multiagent systems. In *Proceedings of the 9th International Conference on Principles on Principles of Knowledge Representation and Reasoning(KR 2004)*, pages 255–266. AAAI Press.

[Boella and van der Torre, 2007] Boella, G. and van der Torre, L. (2007). Norm negotiation in multiagent systems. *International Journal of Cooperative Information Systems*, 16(1):97–122.

[Bratman, 1992] Bratman, M. E. (1992). Practical reasoning and acceptance in context. *Mind*, 101(401):1–15.

[Bulygin, 1992] Bulygin, E. (1992). On norms of competence. *Law and Philosophy*, 11(3):201–216.

[Carmo and Jones, 2002] Carmo, J. and Jones, A. (2002). Deontic logic and contrary-to-duties. In Gabbay, D. and Guenthner, F., editors, *Handbook of Philosophical Logic*, volume 8, pages 265–343. Kluwer Academic Publishers, 2nd edition.

[Castelfranchi, 2003] Castelfranchi, C. (2003). Formalising the informal? Dynamic social order, bottom-up social control, and spontaneous normative relations. *Journal of Applied Logic*, 1(1-2):47–92.

[Chellas, 1980] Chellas, B. F. (1980). *Modal Logic: an Introduction*. Cambridge University Press, Cambridge.

[Clarke, 1994] Clarke, D. (1994). Does acceptance entail belief? *American Philosophical Quarterly*, 31(2):145–155.

[Cohen, 1992] Cohen, L. J. (1992). *An essay on belief and acceptance*. Oxford University Press, New York, USA.

[Cohen and Levesque, 1990] Cohen, P. R. and Levesque, H. J. (1990). Intention is choice with commitment. *Artificial Intelligence*, 42:213–261.

[Coleman, 1990] Coleman, J. (1990). *Foundations of Social Theory*. Harvard University Press, Cambridge.

[Conte et al., 1998] Conte, R., Castelfranchi, C., and Dignum, F. (1998). Autonomous norm acceptance. In *Intelligent Agents V (ATAL'98)*, volume 1555 of *LNCS*, pages 99–112, Berlin. Springer Verlag.

[Dignum and Dignum, 2001] Dignum, V. and Dignum, F. (2001). Modelling agent societies: Coordination frameworks and institutions. In Brazdil, P. and Jorge, A., editors, *Proceedings of the Tenth Portuguese Conference in Artificial Intelligence (EPIA'01)*, volume 2258 of *LNAI*, pages 191–204, Berlin. Springer-Verlag.

[van Ditmarsch et al., 2007] van Ditmarsch, H., van der Hoek, W., and Kooi, B. (2007). *Dynamic Epistemic Logic*, volume 337 of *Synthese Library Series*. Springer.

[Durkheim, 1982] Durkheim, E. (1982). *The rules of Sociological Method*. Free Press, New York. first published in French in 1895.

[Engel, 1998] Engel, P. (1998). Believing, holding true, and accepting. *Philosophical Explorations*, 1(2):140–151.

[Esteva et al., 2001] Esteva, M., Padget, J., and Sierra, C. (2001). Formalizing a language for institutions and norms. In *Intelligent Agents VIII (ATAL'01)*, volume 2333 of *LNAI*, pages 348–366, Berlin. Springer Verlag.

[Fagin et al., 1995] Fagin, R., Halpern, J., Moses, Y., and Vardi, M. (1995). *Reasoning about Knowledge*. MIT Press, Cambridge.

[Gaudou et al., 2008] Gaudou, B., Longin, D., Lorini, E., and Tummolini, L. (2008). Anchoring Institutions in Agents' Attitudes: Towards a Logical Framework for Autonomous MAS. In Padgham, L. and Parkes, D. C., editors, *Proceedings of the Seventh International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS'08)*, pages 728–735. ACM Press.

[Gelati et al., 2004] Gelati, J., Rotolo, A., Sartor, G., and Governatori, G. (2004). Normative autonomy and normative co-ordination: Declarative power, representation, and mandate. *Artificial Intelligence and Law*, 12(1-2):53–81.

[Gerbrandy and Groeneveld, 1997] Gerbrandy, J. and Groeneveld, W. (1997). Reasoning about information change. *Journal of Logic, Language, and Information*, 6:147–196.

[Gilbert, 1987] Gilbert, M. (1987). Modelling collective belief. *Synthese*, 73(1):185–204.

[Gilbert, 1989] Gilbert, M. (1989). *On Social Facts*. Routledge, London and New York.

[Goldblatt, 1992] Goldblatt, R. (1992). *Logics of Time and Computation, 2nd edition*. CSI Lecture Notes, Stanford, California.

[Grossi, 2008] Grossi, D. (2008). Pushing Anderson's envelope: the modal logic of ascription. In *Proceedings of the 9th International Conference on Deontic Logic in Computer Science (DEON'08)*, number 5076 in LNAI, pages 263–277. Springer Verlag.

[Grossi et al., 2006] Grossi, D., Meyer, J.-J. C., and Dignum, F. (2006). Classificatory aspects of counts-as: An analysis in modal logic. *Journal of Logic and Computation*, 16(5):613–643.

[Grossi et al., 2008] Grossi, D., Meyer, J.-J. C., and Dignum, F. (2008). The many faces of counts-as: A formal analysis of constitutive rules. *Journal of Applied Logic*, 6(2):192–217.

[Hakli, 2006] Hakli, P. (2006). Group beliefs and the distinction between belief and acceptance. *Cognitive Systems Research*, 7:286–297.

[Hansen et al., 2007] Hansen, J., Pigozzi, G., and van der Torre, L. (2007). Ten philosophical problems in deontic logic. In Boella, G., van der Torre, L., and Verhagen, H., editors, *Normative Multi-agent Systems*, number 07122 in Dagstuhl Seminar Proceedings. Internationales Begegnungs- und Forschungszentrum fuer Informatik (IBFI), Schloss Dagstuhl, Germany.

[Hart, 1992] Hart, H. L. A. (1992). *The concept of law*. Clarendon Press, Oxford. new edition.

[Herzig et al., 2008] Herzig, A., de Lima, T., and Lorini, E. (2008). What do we accept after an announcement? In Meyer, J.-J. and Broersen, J., editors, *Proceedings of the First Workshop on Knowledge Representation for Agents and Multi-Agent Systems (KRAMAS 2008)*, pages 81–94.

[Hintikka, 1962] Hintikka, J. (1962). *Knowledge and Belief: An Introduction to the Logic of the Two Notions*. Cornell University Press, Ithaca.

[Jones and Sergot, 1996] Jones, A. and Sergot, M. J. (1996). A formal characterization of institutionalised power. *Journal of the IGPL*, 4:429–445.

[Lagerspetz, 2006] Lagerspetz, E. (2006). Institutional facts, performativity and false beliefs. *Cognitive Systems Research*, 7(2-3):298–306.

[Lewis, 1969] Lewis, D. K. (1969). *Convention: a philosophical study*. Harvard University Press, Cambridge.

[Lopez y Lopez et al., 2004] Lopez y Lopez, F., Luck, M., and d'Inverno, M. (2004). Normative agent reasoning in dynamic societies. In *Proceedings of the Third International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS'04)*, pages 732–739. ACM Press.

[Lorini and Longin, 2008] Lorini, E. and Longin, D. (2008). A logical account of institutions: from acceptances to norms via legislators. In Brewka, G. and Lang, J., editors, *Proceedings of the Eleventh International Conference on Principles on Principles of Knowledge Representation and Reasoning (KR 2008)*, pages 38–48. AAAI Press.

[Makinson and van der Torre, 2000] Makinson, D. and van der Torre, L. (2000). Input-output logics. *Journal of Philosophical Logic*, 29:383–408.

[Mantzavinos et al., 2004] Mantzavinos, C., North, D., and Shariq, S. (2004). Learning, institutions, and economic performance. *Perspectives on Politics*, 2:75–84.

[Meyer, 1988] Meyer, J. J. (1988). A different approach to deontic logic: Deontic logic viewed as a variant of dynamic logic. *Notre Dame Journal of Formal Logic*, 29(1):109–136.

[North, 1990] North, D. (1990). *Institutions, Institutional Change, and Economic Performance*. Cambridge University Press, Cambridge.

[Pettit, 2001] Pettit, P. (2001). Deliberative democracy and the discursive dilemma. *Philosophical Issues*, 11:268–99.

[Prakken and Sergot, 1997] Prakken, H. and Sergot, M. (1997). Dyadic deontic logic and contrary-to-duty obligations. In Nute, D., editor, *Defeasible Deontic Logic*, pages 223–262. Synthese Library.

[Rawls, 1955] Rawls, J. (1955). Two concepts of rules. *The Philosophical Review*, 64:3–32.

[Raz, 1975] Raz, J. (1975). *Practical reason and norms*. Hutchinson, London.

[Sahlqvist, 1975] Sahlqvist, H. (1975). Completeness and correspondence in the first and second order semantics for modal logics. In Kanger, S., editor, *Proceedings of the 3rd Scandinavian Logic Symposium*, volume 82 of *Studies in Logic*.

[Searle, 1969] Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language*. Cambridge University Press, New York.

[Searle, 1995] Searle, J. R. (1995). *The Construction of Social Reality*. The Free Press, New York.

[Stalnaker, 1984] Stalnaker, R. (1984). *Inquiry*. MIT Press, Cambridge.

[Tollefsen, 2002] Tollefsen, D. P. (2002). Challenging epistemic individualism. *Protosociology*, 16:86–117.

[Tollefsen, 2003] Tollefsen, D. P. (2003). Rejecting rejectionism. *Protosociology*, 18–19:389–405.

[Tuomela, 1992] Tuomela, R. (1992). Group beliefs. *Synthese*, 91:285–318.

[Tuomela, 2000] Tuomela, R. (2000). Belief versus Acceptance. *Philosophical Explorations*, 2:122–137.

[Tuomela, 2002] Tuomela, R. (2002). *The Philosophy of Social Practices: A Collective Acceptance View*. Cambridge University Press, Cambridge.

[Von Wright, 1963] Von Wright, G. H. (1963). *Norm and Action*. Routledge and Kegan, London.

A Annex: proofs of some theorems

This Annex contains some selected proofs of the theorems presented in the paper.

Proof of Theorem 1

Axiom **K** and rule of inference **Nec** define a minimal normal modal logic. Thus, they do not have an associated semantic constraint. It is a routine task to check that the Axioms **PAccess**, **NAccess**, **Inc**, **Unanim** and **Mon** of the logic $\mathcal{AL}$ correspond to their semantic counterparts **S.1**-**S.5** over $\mathcal{AL}$ models. In particular, the following correspondences exist between the axioms of the logic $\mathcal{AL}$ and the semantic constraints over $\mathcal{AL}$ frames.

- Axioms **PAccess** corresponds to the constraint **S.1**.
- Axiom **NAccess** corresponds to the constraint **S.2**.
- Axiom **Inc** corresponds to the constraint **S.3**.
- Axiom **Unanim** corresponds to the constraint **S.4**.
- Axiom **Mon** corresponds to the constraint **S.5**.

It is a routine, too, to check that all of axioms of the logic $\mathcal{AL}$ are in the Sahlqvist class, for which a general completeness result exists.
(See [Sahlqvist, 1975, Blackburn et al., 2001].)

Proof of Theorem 2

For notational convenience, we will use the following abbreviation in the proof:

$$\widehat{\mathcal{A}}_{C:x}\varphi \stackrel{def}{=} \neg\mathcal{A}_{C:x}\neg\varphi$$

We have to prove that if φ is $\mathcal{AL}$ *satisfiable* then it is satisfiable in a finite $\mathcal{AL}$ model.

Suppose that $\mathcal{M} = \langle W, \mathscr{A}, \mathscr{V}\rangle$ is a $\mathcal{AL}$ model which satisfies φ. Our aim is to build a finite $\mathcal{AL}$ model which satisfies φ. To do this, we use a filtration method [Blackburn et al., 2001, Goldblatt, 1992].

Let us introduce the following definition.

Definition 1. *A set of formulas Σ is closed under subformulas (cus) if for all formulas φ, φ': if $\varphi \vee \varphi' \in \Sigma$ then so are φ and φ'; if $\neg\varphi' \in \Sigma$ then so is φ; for any $x \in INST$ and $C \in 2^{AGT\star}$ if $\mathcal{A}_{C:x}\varphi \in \Sigma$ then $\varphi \in \Sigma$.*

Let us now consider an arbitrary finite set of formulas Σ_φ which is *closed under subformulas* and which contains φ. From Σ_φ we define the set Σ^+_φ as follows.

Σ^+_φ is defined as the smallest superset of Σ_φ such that:

1. for all $x, y \in INST$ and $C, B \in 2^{AGT\star}$, if $\mathcal{A}_{C:x}\varphi \in \Sigma^+_\varphi$ then $\mathcal{A}_{B:y}\varphi \in \Sigma^+_\varphi$;

2. for all $x \in INST$ and $C \in 2^{AGT\star}$, if $\mathcal{A}_{C:x}\varphi \in \Sigma_\varphi^+$ then $\neg\mathcal{A}_{C:x}\varphi \in \Sigma_\varphi^+$;

3. for all $x \in INST$ and $C \in 2^{AGT\star}$, $\mathcal{A}_{C:x}\bot \in \Sigma_\varphi^+$;

4. $\bot \in \Sigma_\varphi^+$.

The following proposition follows straightforwardly due to the fact that the sets AGT and $INST$ are supposed to be finite.

Proposition 3. *Σ_φ^+ is finite and closed under subformulas.*

We define the relation $\leftrightsquigarrow$ between the worlds in W of the model $\mathcal{M}$. For every two worlds $w, v \in W$:

- $w \leftrightsquigarrow v$ iff for all $\varphi \in \Sigma_\varphi^+$, $\mathcal{M}, w \models \varphi$ iff $\mathcal{M}, v \models \varphi$.

For every world $w \in W$, we note $|w|$ the equivalence class of world w of $\mathcal{M}$ with respect to $\leftrightsquigarrow$. Moreover, let $W_{\Sigma_\varphi^+} = \{|w| \mid w \in W\}$.

Now, we have to build a filtrated model $\mathcal{M}^f = \langle W^f, \mathscr{A}^f, \mathscr{V}^f \rangle$ of the model $\mathcal{M}$.

Definition 2. *We define $\mathcal{M}^f$ as follows.*

A. $W^f = W_{\Sigma_\varphi^+}$;

B. *for every $B \in 2^{AGT\star}$ and $x \in INST$, $|v| \in \mathscr{A}^f_{B:x}(|w|)$ if and only if:*

1. $\forall \mathcal{A}_{B:x}\varphi \in \Sigma_\varphi^+$, *if* $\mathcal{M}, w \models \mathcal{A}_{B:x}\varphi$ *then* $\mathcal{M}, v \models \varphi$;

2. $\forall y \in INST$ *and* $\forall C \in 2^{AGT\star}$, *if* $B \subseteq C$ *then:*
$\forall \mathcal{A}_{C:y}\varphi \in \Sigma_\varphi^+$, *if* $\mathcal{M}, w \models \mathcal{A}_{C:y}\varphi$ *then* $\mathcal{M}, v \models \mathcal{A}_{C:y}\varphi$;

3. $\forall y \in INST$ *and* $\forall C \in 2^{AGT\star}$, *if* $B \subseteq C$ *then:*
$\forall \widehat{\mathcal{A}}_{C:y}\varphi \in \Sigma_\varphi^+$, *if* $\mathcal{M}, w \models \widehat{\mathcal{A}}_{C:y}\varphi$ *then* $\mathcal{M}, v \models \widehat{\mathcal{A}}_{C:y}\varphi$;

4. $\forall C \in 2^{AGT\star}$, *if* $B \subseteq C$ *then:*
$\forall \mathcal{A}_{C:x}\varphi, \widehat{\mathcal{A}}_{C:x}\top \in \Sigma_\varphi^+$, *if* $\mathcal{M}, w \models \widehat{\mathcal{A}}_{C:x}\top \wedge \mathcal{A}_{C:x}\varphi$ *then* $\mathcal{M}, v \models \varphi$;

5. $\exists i \in B$ *such that* $\forall \mathcal{A}_{i:x}\varphi \in \Sigma_\varphi^+$, *if* $\mathcal{M}, v \models \mathcal{A}_{i:x}\varphi$ *then* $\mathcal{M}, v \models \varphi$.

C. $\mathscr{V}^f(p) = \{|w| \mid \mathcal{M}, w \models p\}$, *for all propositional atoms in* Σ_φ^+.

It is straightforward to prove that the model $\mathcal{M}^f$ is indeed a filtration of $\mathcal{M}$ through Σ_φ^+.

Lemma 1. *$\mathcal{M}^f$ is a filtration of $\mathcal{M}$ through Σ_φ^+.*

The next step consists in proving that $\mathcal{M}^f$ is a $\mathcal{AL}$ model.

Lemma 2. *$\mathcal{M}^f$ is a $\mathcal{AL}$ model.*

Proof. We have to prove that the model $\mathcal{M}^f$ satisfies the five semantic constraints **S.1**-**S.5** over $\mathcal{AL}$ models.

Let us start with constraint **S.1**. We have to prove that the following condition holds in $\mathcal{M}^f$ for any $x, y \in INST$ and $C, B \in 2^{AGT\star}$ such that $B \subseteq C$:

- if $|w'| \in \mathscr{A}^f_{B:x}(|w|)$ and $|w''| \in \mathscr{A}^f_{C:y}(|w'|)$ then $|w''| \in \mathscr{A}^f_{C:y}(|w|)$.

Suppose $|w'| \in \mathscr{A}^f_{B:x}(|w|)$ and $w'' \in \mathscr{A}^f_{C:y}(|w'|)$, where $B \subseteq C$. We have to prove that $|w''| \in \mathscr{A}^f_{C:y}(|w|)$. By Definition 2, the latter is equivalent to:

1. $\forall \mathcal{A}_{C:y}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w \models \mathcal{A}_{C:y}\varphi$ then $\mathcal{M}, w'' \models \varphi$;

2. $\forall z \in INST$ and $\forall D \in 2^{AGT\star}$, if $C \subseteq D$ then:
$\forall \mathcal{A}_{D:z}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w \models \mathcal{A}_{D:z}\varphi$ then $\mathcal{M}, w'' \models \mathcal{A}_{D:z}\varphi$;

3. $\forall z \in INST$ and $\forall D \in 2^{AGT\star}$, if $C \subseteq D$ then:
$\forall \widehat{\mathcal{A}}_{D:z}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w \models \widehat{\mathcal{A}}_{D:z}\varphi$ then $\mathcal{M}, w'' \models \widehat{\mathcal{A}}_{D:z}\varphi$;

4. $\forall D \in 2^{AGT\star}$, if $C \subseteq D$ then:
$\forall \mathcal{A}_{D:y}\varphi, \widehat{\mathcal{A}}_{D:y}\top \in \Sigma^+_\varphi$, if $\mathcal{M}, w \models \widehat{\mathcal{A}}_{D:y}\top \wedge \mathcal{A}_{D:y}\varphi$ then $\mathcal{M}, w'' \models \varphi$;

5. $\exists i \in C$ such that $\forall \mathcal{A}_{i:y}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w'' \models \mathcal{A}_{i:y}\varphi$ then $\mathcal{M}, w'' \models \varphi$.

So, to prove **S.1** we just need to prove that the previous items **1**-**5** are consequences of $|w'| \in \mathscr{A}^f_{B:x}(|w|)$ and $|w''| \in \mathscr{A}^f_{C:y}(|w'|)$ when $B \subseteq C$.

Item 1. Suppose $\mathcal{A}_{C:y}\varphi \in \Sigma^+_\varphi$ and $\mathcal{M}, w \models \mathcal{A}_{C:y}\varphi$. As $B \subseteq C$ and $|w'| \in \mathscr{A}^f_{B:x}(|w|)$, it follows that $\mathcal{M}, w'' \models \varphi$.

Item 2. Take an arbitrary D such that $C \subseteq D$ and an arbitrary $z \in INST$. As $B \subseteq C$, we have $B \subseteq D$. Suppose $\mathcal{A}_{D:z}\varphi \in \Sigma^+_\varphi$ and $\mathcal{M}, w \models \mathcal{A}_{D:z}\varphi$. As $|w'| \in \mathscr{A}^f_{B:x}(|w|)$, it follows that $\mathcal{M}, w' \models \mathcal{A}_{D:z}\varphi$. As $w'' \in \mathscr{A}^f_{C:y}(|w'|)$, we conclude that $\mathcal{M}, w'' \models \mathcal{A}_{D:z}\varphi$.

Item 3. Take an arbitrary D such that $C \subseteq D$ and an arbitrary $z \in INST$. As $B \subseteq C$, we have $B \subseteq D$. Suppose $\widehat{\mathcal{A}}_{D:z}\varphi \in \Sigma^+_\varphi$ and $\mathcal{M}, w \models \widehat{\mathcal{A}}_{D:z}\varphi$. As $|w'| \in \mathscr{A}^f_{B:x}(|w|)$, it follows that $\mathcal{M}, w' \models \widehat{\mathcal{A}}_{D:z}\varphi$. As $w'' \in \mathscr{A}^f_{C:y}(|w'|)$, we conclude that $\mathcal{M}, w'' \models \widehat{\mathcal{A}}_{D:z}\varphi$.

Item 4. Take an arbitrary D such that $C \subseteq D$. As $B \subseteq C$, we have $B \subseteq D$. Suppose $\mathcal{A}_{D:y}\varphi, \widehat{\mathcal{A}}_{D:y}\top \in \Sigma^+_\varphi$ and $\mathcal{M}, w \models \mathcal{A}_{D:y}\varphi \wedge \widehat{\mathcal{A}}_{D:y}\top$. As $|w'| \in \mathscr{A}^f_{B:x}(|w|)$, it follows that $\mathcal{M}, w' \models \mathcal{A}_{D:y}\varphi \wedge \widehat{\mathcal{A}}_{D:y}\top$. As $w'' \in \mathscr{A}^f_{C:y}(|w'|)$, we conclude that $\mathcal{M}, w'' \models \varphi$.

Item 5. This item follows straightforwardly from the fact $w'' \in \mathscr{A}^f_{C:y}(|w'|)$.

This proves that **S.1** holds.

Let us now consider constraint **S.2**. We have to prove that the following condition holds in $\mathcal{M}^f$ for any $x, y \in INST$ and $C, B \in 2^{AGT\star}$ such that $B \subseteq C$:

- if $|w'| \in \mathscr{A}^f_{B:x}(|w|)$ and $w'' \in \mathscr{A}^f_{C:y}(|w|)$ then $w'' \in \mathscr{A}^f_{C:y}(|w'|)$.

Suppose $|w'| \in \mathscr{A}^f_{B:x}(|w|)$ and $w'' \in \mathscr{A}^f_{C:y}(|w|)$, where $B \subseteq C$. We have to prove that $|w''| \in \mathscr{A}^f_{C:y}(|w'|)$. By Definition 2, the latter is equivalent to:

1. $\forall \mathcal{A}_{C:y}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w' \models \mathcal{A}_{C:y}\varphi$ then $\mathcal{M}, w'' \models \varphi$;

2. $\forall z \in INST$ and $\forall D \in 2^{AGT\star}$, if $C \subseteq D$ then:
$\forall \mathcal{A}_{D:z}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w' \models \mathcal{A}_{D:z}\varphi$ then $\mathcal{M}, w'' \models \mathcal{A}_{D:z}\varphi$;

3. $\forall z \in INST$ and $\forall D \in 2^{AGT\star}$, if $C \subseteq D$ then:
$\forall \widehat{\mathcal{A}}_{D:z}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w' \models \widehat{\mathcal{A}}_{D:z}\varphi$ then $\mathcal{M}, w'' \models \widehat{\mathcal{A}}_{D:z}\varphi$;

4. $\forall D \in 2^{AGT\star}$, if $C \subseteq D$ then:
$\forall \mathcal{A}_{D:y}\varphi, \widehat{\mathcal{A}}_{D:y}\top \in \Sigma^+_\varphi$, if $\mathcal{M}, w' \models \widehat{\mathcal{A}}_{D:y}\top \wedge \mathcal{A}_{D:y}\varphi$ then $\mathcal{M}, w'' \models \varphi$;

5. $\exists i \in C$ such that $\forall \mathcal{A}_{i:y}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w'' \models \mathcal{A}_{i:y}\varphi$ then $\mathcal{M}, w'' \models \varphi$.

So, to prove **S.2** we just need to prove that items **1**-**5** are consequences of $|w'| \in \mathscr{A}^f_{B:x}(|w|)$ and $w'' \in \mathscr{A}^f_{C:y}(|w|)$.

Item 1. Suppose $\mathcal{A}_{C:y}\varphi \in \Sigma^+_\varphi$ and $\mathcal{M}, w' \models \mathcal{A}_{C:y}\varphi$. By construction of Σ^+_φ we have $\widehat{\mathcal{A}}_{C:y}\neg\varphi \in \Sigma^+_\varphi$. As $|w'| \in \mathscr{A}^f_{B:x}(|w|)$ and $B \subseteq C$, it follows that $\mathcal{M}, w \models \mathcal{A}_{C:y}\varphi$. As $|w''| \in \mathscr{A}^f_{C:y}(|w|)$, we conclude that $\mathcal{M}, w'' \models \varphi$.

Item 2. Take an arbitrary D such that $C \subseteq D$ and an arbitrary $z \in INST$. As $B \subseteq C$, we have $B \subseteq D$. Suppose $\mathcal{A}_{D:z}\varphi \in \Sigma^+_\varphi$ and $\mathcal{M}, w' \models \mathcal{A}_{D:z}\varphi$. By construction of Σ^+_φ we have $\widehat{\mathcal{A}}_{D:z}\neg\varphi \in \Sigma^+_\varphi$. As $|w'| \in \mathscr{A}^f_{B:x}(|w|)$ and $B \subseteq D$, it follows that $\mathcal{M}, w \models \mathcal{A}_{D:z}\varphi$. As $|w''| \in \mathscr{A}^f_{C:y}(|w|)$, we conclude that $\mathcal{M}, w'' \models \mathcal{A}_{D:z}\varphi$.

Item 3. Take an arbitrary D such that $C \subseteq D$ and an arbitrary $z \in INST$. As $B \subseteq C$, we have $B \subseteq D$. Suppose $\widehat{\mathcal{A}}_{D:z}\varphi \in \Sigma^+_\varphi$ and $\mathcal{M}, w' \models \widehat{\mathcal{A}}_{D:z}\varphi$. By construction of Σ^+_φ we have $\mathcal{A}_{D:z}\neg\varphi \in \Sigma^+_\varphi$. As $|w'| \in \mathscr{A}^f_{B:x}(|w|)$ and $C \subseteq D$, it follows that $\mathcal{M}, w \models \widehat{\mathcal{A}}_{D:z}\varphi$. As $|w''| \in \mathscr{A}^f_{C:y}(|w|)$, we conclude that $\mathcal{M}, w'' \models \widehat{\mathcal{A}}_{D:z}\varphi$.

Item 4. Take an arbitrary D such that $C \subseteq D$. As $B \subseteq C$, we have $B \subseteq D$. Suppose $\mathcal{A}_{D:y}\varphi, \widehat{\mathcal{A}}_{D:y}\top \in \Sigma^+_\varphi$ and $\mathcal{M}, w' \models \mathcal{A}_{D:y}\varphi \wedge \widehat{\mathcal{A}}_{D:y}\top$. By construction of Σ^+_φ we have $\mathcal{A}_{D:y}\bot, \widehat{\mathcal{A}}_{D:y}\neg\varphi \in \Sigma^+_\varphi$. As $|w'| \in \mathscr{A}^f_{B:x}(|w|)$ and $C \subseteq D$, it follows that $\mathcal{M}, w \models \mathcal{A}_{D:y}\varphi \wedge \widehat{\mathcal{A}}_{D:y}\top$. As $|w''| \in \mathscr{A}^f_{C:y}(|w|)$, we conclude that $\mathcal{M}, w'' \models \varphi$.

Item 5. This item follows straightforwardly from the fact $w'' \in \mathscr{A}^f_{C:y}(|w|)$.

This proves that **S.2** holds.

As a next step we have to prove the model $\mathcal{M}^f$ satisfies the semantic condition **S.3**. That is, we have to prove that for any $x \in INST$ and $C, B \in 2^{AGT\star}$ such that $B \subseteq C$:

- if $\mathscr{A}^f_{C:x}(|w|) \neq \emptyset$ then $\mathscr{A}^f_{B:x}(|w|) \subseteq \mathscr{A}^f_{C:x}(|w|)$.

The following proposition is needed to prove that $\mathcal{M}^f$ satisfies the condition **S.3**.

Proposition 4. *For every $x \in INST$ and $C \in 2^{AGT\star}$, if $\mathscr{A}^f_{C:x}(|w|) \neq \emptyset$ then $\exists w \in |w|$ such that $\mathcal{M}, w \models \widehat{\mathcal{A}}_{C:x}\top$.*

Proof. Let us suppose that $\mathscr{A}^f_{C:x}(|w|) \neq \emptyset$, and $\mathcal{M}, w \models \mathcal{A}_{C:x}\bot$ for all $w \in |w|$. We are going to show that the two facts are inconsistent.

Condition $\mathscr{A}^f_{C:x}(|w|) \neq \emptyset$ implies that $\exists |w'| \in W^f$ such that: if $\mathcal{A}_{C:x}\varphi \in \Sigma^+_\varphi$ then, if $\mathcal{M}, w \models \mathcal{A}_{C:x}\varphi$ then $\mathcal{M}, w' \models \varphi$. As we have $\mathcal{A}_{C:x}\bot \in \Sigma^+_\varphi$ (by construction of Σ^+_φ) and we have supposed $\mathcal{M}, w \models \mathcal{A}_{C:x}\bot$, we can infer that $\mathcal{M}, w' \models \bot$. $\square$

Let us now prove that $\mathcal{M}^f$ satisfies the condition **S.3**. Consider an arbitrary $x \in INST$ and $C, B \in 2^{AGT\star}$ such that $B \subseteq C$. Suppose that $\mathscr{A}^f_{C:x}(|w|) \neq \emptyset$ and $w' \in \mathscr{A}^f_{B:x}(|w|)$. We have to prove that $w' \in \mathscr{A}^f_{C:x}(|w|)$. By Definition 2, the latter is equivalent to:

1. $\forall \mathcal{A}_{C:x}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w \models \mathcal{A}_{C:x}\varphi$ then $\mathcal{M}, w' \models \varphi$;

2. $\forall z \in INST$ and $\forall D \in 2^{AGT\star}$, if $C \subseteq D$ then:
$\forall \mathcal{A}_{D:z}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w \models \mathcal{A}_{D:z}\varphi$ then $\mathcal{M}, w' \models \mathcal{A}_{D:z}\varphi$;

3. $\forall z \in INST$ and $\forall D \in 2^{AGT\star}$, if $C \subseteq D$ then:
$\forall \widehat{\mathcal{A}}_{D:z}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w \models \widehat{\mathcal{A}}_{D:z}\varphi$ then $\mathcal{M}, w' \models \widehat{\mathcal{A}}_{D:z}\varphi$;

4. $\forall D \in 2^{AGT\star}$, if $C \subseteq D$ then:
$\forall \mathcal{A}_{D:x}\varphi, \widehat{\mathcal{A}}_{D:x}\top \in \Sigma^+_\varphi$, if $\mathcal{M}, w \models \widehat{\mathcal{A}}_{D:x}\top \wedge \mathcal{A}_{D:x}\varphi$ then $\mathcal{M}, w' \models \varphi$;

5. $\exists i \in C$ such that $\forall \mathcal{A}_{i:x}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w' \models \mathcal{A}_{i:x}\varphi$ then $\mathcal{M}, w' \models \varphi$.

So, to prove that $\mathcal{M}^f$ satisfies the condition **S.3** we just need to prove that items **1**-**5** are consequences of $\mathscr{A}^f_{C:x}(|w|) \neq \emptyset$ and $|w'| \in \mathscr{A}^f_{B:x}(|w|)$.

Item 1. Suppose $\mathcal{A}_{C:x}\varphi \in \Sigma^+_\varphi$ and $\mathcal{M}, w \models \mathcal{A}_{C:x}\varphi$. By construction of Σ^+_φ, we have $\widehat{\mathcal{A}}_{C:x}\top \in \Sigma^+_\varphi$. From $\mathscr{A}^f_{C:x}(|w|) \neq \emptyset$ it follows that $\exists w \in |w|$ such that $\mathcal{M}, w \models \widehat{\mathcal{A}}_{C:x}\top$ (by Proposition 4). Thus, by definition of $|w|$, we can conclude that $\forall w \in |w|$ it holds that $\mathcal{M}, w \models \widehat{\mathcal{A}}_{C:x}\top$. Then, in particular, $\mathcal{M}, w \models \widehat{\mathcal{A}}_{C:x}\top$. As $\mathcal{M}, w \models \mathcal{A}_{C:x}\varphi$ and $B \subseteq C$, from the latter it follows that $\mathcal{M}, w \models \mathcal{A}_{B:x}\varphi$ (by Axiom **Inc** of the logic $\mathcal{AL}$). As $|w'| \in \mathscr{A}^f_{B:x}(|w|)$ and $\mathcal{A}_{B:x}\varphi \in \Sigma^+_\varphi$ (from $\mathcal{A}_{C:x}\varphi \in \Sigma^+_\varphi$, by construction of Σ^+_φ), from the latter we conclude $\mathcal{M}, w' \models \varphi$.

Item 2. Take an arbitrary D such that $C \subseteq D$ and an arbitrary $z \in INST$. As $B \subseteq C$, we have $B \subseteq D$. Moreover, suppose $\mathcal{A}_{D:z}\varphi \in \Sigma^+_\varphi$ and $\mathcal{M}, w \models \mathcal{A}_{D:z}\varphi$. As $|w'| \in \mathscr{A}^f_{B:x}(|w|)$, we conclude that $\mathcal{M}, w' \models \mathcal{A}_{D:z}\varphi$.

Item 3. Take an arbitrary D such that $C \subseteq D$ and an arbitrary $z \in INST$. As $B \subseteq C$, we have $B \subseteq D$. Moreover, suppose $\widehat{\mathcal{A}}_{D:z}\varphi \in \Sigma^+_\varphi$ and $\mathcal{M}, w \models \widehat{\mathcal{A}}_{D:z}\varphi$. As $|w'| \in \mathscr{A}^f_{B:x}(|w|)$, we conclude that $\mathcal{M}, w' \models \widehat{\mathcal{A}}_{D:z}\varphi$.

Item 4. Take an arbitrary D such that $C \subseteq D$. As $B \subseteq C$, we have $B \subseteq D$. Moreover, suppose $\mathcal{A}_{D:x}\varphi, \widehat{\mathcal{A}}_{D:x}\top \in \Sigma^+_\varphi$ and $\mathcal{M}, w \models \mathcal{A}_{D:x}\varphi \wedge \widehat{\mathcal{A}}_{D:x}\top$. As $|w'| \in \mathscr{A}^f_{B:x}(|w|)$, we conclude that $\mathcal{M}, w' \models \varphi$.

Item 5. From $w' \in \mathscr{A}^f_{B:x}(|w|)$, it follows that $\exists i \in B$ such that $\forall \mathcal{A}_{i:x}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w' \models \mathcal{A}_{i:x}\varphi$ then $\mathcal{M}, w' \models \varphi$. As $B \subseteq C$, the latter implies that $\exists i \in C$ such that $\forall \mathcal{A}_{i:x}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w' \models \mathcal{A}_{i:x}\varphi$ then $\mathcal{M}, w' \models \varphi$.

This proves that **S.3** holds.

Now, we prove that the model $\mathcal{M}^f$ satisfies the semantic condition **S.4**. That is, we prove that for any $x \in INST$ and $C \in 2^{AGT\star}$:

- if $|w'| \in \mathscr{A}^f_{C:x}(|w|)$ then $|w'| \in \bigcup_{i \in C} \mathscr{A}_{i:x}(|w'|)$.

Suppose $|w'| \in \mathscr{A}^f_{C:x}(|w|)$. We have to prove that $|w'| \in \bigcup_{i \in C} \mathscr{A}_{i:x}(|w'|)$. By Definition 2, the latter is equivalent to the fact that $\exists i \in C$ such that:

1. $\forall \mathcal{A}_{i:x}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w' \models \mathcal{A}_{i:x}\varphi$ then $\mathcal{M}, w' \models \varphi$;

2. $\forall z \in INST$ and $\forall D \in 2^{AGT\star}$, if $i \in D$ then:
$\forall \mathcal{A}_{D:z}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w' \models \mathcal{A}_{D:z}\varphi$ then $\mathcal{M}, w' \models \mathcal{A}_{D:z}\varphi$;

3. $\forall z \in INST$ and $\forall D \in 2^{AGT\star}$, if $i \in D$ then:
$\forall \widehat{\mathcal{A}}_{D:z}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w' \models \widehat{\mathcal{A}}_{D:z}\varphi$ then $\mathcal{M}, w' \models \widehat{\mathcal{A}}_{D:z}\varphi$;

4. $\forall D \in 2^{AGT\star}$, if $i \in D$ then:
$\forall \mathcal{A}_{D:x}\varphi, \widehat{\mathcal{A}}_{D:x}\top \in \Sigma^+_\varphi$, if $\mathcal{M}, w' \models \widehat{\mathcal{A}}_{D:x}\top \wedge \mathcal{A}_{D:x}\varphi$ then $\mathcal{M}, w' \models \varphi$;

5. $\forall \mathcal{A}_{i:x}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w' \models \mathcal{A}_{i:x}\varphi$ then $\mathcal{M}, w' \models \varphi$.

Thus, we have to suppose $|w'| \in \mathscr{A}^f_{C:x}(|w|)$ and prove that $\exists i \in C$ which satisfies items **1**-**5**. Items **2** and **3** trivially hold for all $\exists i \in C$. Moreover, items **1** and **5** are the same condition. Therefore, we just need to prove that $|w'| \in \mathscr{A}^f_{C:x}(|w|)$ implies that $\exists i \in C$ which satisfies items **1** and **4**.

From $|w'| \in \mathscr{A}^f_{C:x}(|w|)$, we can infer that $\exists i \in C$ such that $\forall \mathcal{A}_{i:x}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w' \models \mathcal{A}_{i:x}\varphi$ then $\mathcal{M}, w' \models \varphi$.

By Axiom **Inc** of the logic $\mathcal{AL}$ and by construction of Σ^+_φ the following property holds for all $i \in C$. For all $D \in 2^{AGT\star}$, if $i \in D$ then: $\forall \mathcal{A}_{D:x}\varphi, \widehat{\mathcal{A}}_{D:x}\top \in \Sigma^+_\varphi$, if $\mathcal{M}, w' \models \widehat{\mathcal{A}}_{D:x}\top \wedge \mathcal{A}_{D:x}\varphi$ then $\mathcal{M}, w' \models \mathcal{A}_{i:x}\varphi$ and $\mathcal{A}_{i:x}\varphi \in \Sigma^+_\varphi$.

From the previous two facts, we conclude that $\exists i \in C$ such that: $\forall \mathcal{A}_{i:x}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w' \models \mathcal{A}_{i:x}\varphi$ then $\mathcal{M}, w' \models \varphi$; $\forall D \in 2^{AGT\star}$, if $i \in D$ then: $\forall \mathcal{A}_{D:x}\varphi, \widehat{\mathcal{A}}_{D:x}\top \in \Sigma^+_\varphi$, if $\mathcal{M}, w' \models \widehat{\mathcal{A}}_{D:x}\top \wedge \mathcal{A}_{D:x}\varphi$ then $\mathcal{M}, w' \models \varphi$.

This proves that **S.4** holds.

It remains to be proved that the model $\mathcal{M}^f$ satisfies the semantic condition **S.5**. That is, we have to prove that for any $x \in INST$ and $C, B \in 2^{AGT\star}$ such that $B \subseteq C$:

- if $\mathscr{A}^f_{C:x}(|w|) \neq \emptyset$ then $\mathscr{A}^f_{B:x}(|w|) \neq \emptyset$.

In order to prove this, we prove first that $\mathscr{A}^f_{C:x}(|w|) \neq \emptyset$ implies $\mathcal{M}, w \models \widehat{\mathcal{A}}_{B:x}\top$, when $B \subseteq C$.

Let us suppose that $\mathscr{A}^f_{C:x}(|w|) \neq \emptyset$ and $\mathcal{M}, w \models \mathcal{A}_{B:x}\bot$ with $B \subseteq C$. We show that these facts are inconsistent.

From $\mathcal{M}, w \models \mathcal{A}_{B:x}\bot$ we infer $\mathcal{M}, w \models \mathcal{A}_{C:x}\bot$ (by Axiom **Mon** of the logic $\mathcal{AL}$ and the fact that $B \subseteq C$). From Definition 2 and $\mathscr{A}^f_{C:x}(|w|) \neq \emptyset$, we can infer that $\exists |w'|$ such that $\forall \mathcal{A}_{C:x}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w \models \mathcal{A}_{C:x}\varphi$ then $\mathcal{M}, w' \models \varphi$. By construction of Σ^+_φ we have that $\mathcal{A}_{C:x}\bot \in \Sigma^+_\varphi$. Thus, as we have $\mathcal{M}, w \models \mathcal{A}_{C:x}\bot$, we conclude that $\exists |w'|$ such that $\mathcal{M}, w' \models \bot$.

This proves that $\mathscr{A}^f_{C:x}(|w|) \neq \emptyset$ implies $\mathcal{M}, w \models \widehat{\mathcal{A}}_{B:x}\top$, when $B \subseteq C$.

Now, we have to show that $\mathcal{M}, w \models \widehat{\mathcal{A}}_{B:x}\top$ implies $\mathscr{A}^f_{B:x}(|w|) \neq \emptyset$.

$\mathcal{M}, w \models \widehat{\mathcal{A}}_{B:x}\top$ implies that $\exists w'$ such that $w' \in \mathscr{A}_{B:x}(w)$. As $\mathcal{M}^f$ is a filtration of $\mathcal{M}$ (Lemma 1), from the latter we conclude that $\exists |w'|$ such that $|w'| \in \mathscr{A}^f_{B:x}(|w|)$.

This proves that **S.5** holds.

□

Lemma 3. *The model $\mathcal{M}^f$ contains at most 2^n worlds where n denotes the size of Σ^+_φ.*

Proof. From Lemma 1 and Proposition 2.38 given in [Blackburn et al., 2001, p. 79]. □

Lemma 4. *$\mathcal{M}^f$ is a finite model.*

Proof. From Lemma 3 and Proposition 3. □

Lemma 5. *Formula φ is satisfiable in $\mathcal{M}^f$.*

Proof. From Lemma 1 and Proposition 3, the fact that φ is satisfiable in $\mathcal{M}$, the fact that $\varphi \in \Sigma^+_\varphi$ and the Filtration Theorem given in [Blackburn et al., 2001, p. 79]. □

Lemma 6. *The logic $\mathcal{AL}$ has the finite model property.*

Proof. We have started with an arbitrary formula φ which is satisfiable in a $\mathcal{AL}$ model $\mathcal{M}$. We have built a model $\mathcal{M}^f$ and proved that $\mathcal{M}^f$ is a finite $\mathcal{AL}$ model (Lemma 4). Finally, we have proved that φ is satisfiable in $\mathcal{M}^f$ (Lemma 5). Thus, we can conclude that for every formula φ, if φ is $\mathcal{AL}$ satisfiable then, φ is satisfiable in a finite $\mathcal{AL}$ model. □

Theorem 2 is a direct consequence of Lemma 6.

Proof of Theorem 14

As for the logic $\mathcal{AL}$, it is a routine to prove soundness, whereas completeness is again obtained by Sahlqvist completeness theorem. Indeed, all axioms of $\mathcal{AL}^+$ are in the Sahlqvist class, for which a general completeness result exists [Sahlqvist, 1975, Blackburn et al., 2001].

Proof of Theorem 15

In order to prove Theorem 15, it is sufficient to prove that if $INST = CXT$ and φ is a formula of the $\mathcal{GMD}$ logic then: if φ is a theorem of $\mathcal{GMD}$ then $tr(\varphi)$ is a theorem of $\mathcal{AL}^+$ and, if φ is $\mathcal{GMD}$ satisfiable then $tr(\varphi)$ is $\mathcal{AL}^+$ satisfiable.

Proposition 5. *Suppose that $INST = CXT$ and φ is a formula of the logic $\mathcal{GMD}$ then: if $\vdash_{\mathcal{GMD}} \varphi$ then $\vdash_{\mathcal{AL}^+} tr(\varphi)$.*

Proof. We only need to prove that the translations of the axioms of the $\mathcal{GMD}$ logic are theorems of $\mathcal{AL}^+$ and that the translated rules of inference of $\mathcal{GMD}$ preserves validity.

It is straightforward to show that the translation of the rules of inference $\mathbf{Nec}_{[\![x]\!]}$, $\mathbf{Nec}_{[\![Univ]\!]}$ and **MP** preserve validity. As the $\mathcal{AL}^+$ operators $[x]$ and $[Univ]$ are normal, it is a routine to verify that the translation of the $\mathcal{GMD}$ Axioms $\mathbf{K}_{[\![x]\!]}$ and $\mathbf{K}_{[\![Univ]\!]}$ are

theorems of $\mathcal{AL}^+$. Furthermore, by the definitions of $[x]\,\varphi$ and $[Univ]\,\varphi$, it is just trivial to prove that that the translation of the $\mathcal{GMD}$ Axiom $\subseteq_{[\![Univ]\!],[\![x]\!]}$ is a theorem of $\mathcal{AL}^+$. The translation of the $\mathcal{GMD}$ Axiom $\mathbf{T}_{[\![Univ]\!]}$ is a theorem of $\mathcal{AL}^+$ as well. Indeed, this corresponds to the Axiom $\mathbf{T}_{Univ}$ of the logic $\mathcal{AL}^+$. By Axioms **PAccess**$^+$ and **NAccess**$^+$ we can prove that the translations of the $\mathcal{GMD}$ Axioms $\mathbf{4}_{[\![x]\!],[\![y]\!]}$ and $\mathbf{5}_{[\![x]\!],[\![y]\!]}$ are theorems of $\mathcal{AL}^+$. By the same principles, we can prove that the translations of the $\mathcal{GMD}$ Axioms $\mathbf{4}_{[\![Univ]\!]}$ and $\mathbf{5}_{[\![Univ]\!]}$ are theorems of $\mathcal{AL}^+$. □

Proposition 6. *Suppose that $INST = CXT$ and φ is a formula of the logic $\mathcal{GMD}$ then: if φ is $\mathcal{GMD}$ satisfiable then $tr(\varphi)$ is $\mathcal{AL}^+$ satisfiable.*

Proof. Suppose that φ is $\mathcal{GMD}$ satisfiable. Thus, there exists a $\mathcal{GMD}$ model $\mathcal{M}^{\mathcal{GMD}} = \langle S, \{S_x\}_{x \in CXT_0}, \pi \rangle$ which satisfies φ. We prove that we can build a $\mathcal{AL}^+$ model $\mathcal{M}$ which satisfies the same formulas as $\mathcal{M}^{\mathcal{GMD}}$.

As we have supposed $INST = CXT$, the $\mathcal{AL}^+$ model $\mathcal{M}$ associated with the $\mathcal{GMD}$ model $\mathcal{M}^{\mathcal{GMD}}$ can be defined as follows.

- $W = S$;
- $\forall w \in W, \forall x \in CXT_0, \forall C \in 2^{AGT\star}$, $\mathscr{A}_{C:x}(w) = S_x$;
- $\forall w \in W, \forall C \in 2^{AGT\star}$, $\mathscr{A}_{C:Univ}(w) = S$;
- $\forall w \in W, \forall p \in ATM, w \in \pi(p)$ if and only if $w \in \mathscr{V}(p)$.

It is a routine to verify that the previous conditions ensure that the model $\mathcal{M}$ is indeed a $\mathcal{AL}^+$ model. By structural induction on φ, it is also a routine to prove that the previous $\mathcal{AL}^+$ model satisfies the same formulas as the $\mathcal{GMD}$ model it is associated. That is, $\mathcal{M}^{\mathcal{GMD}}, w \models \varphi$ if and only if $\mathcal{M}, w \models tr(\varphi)$.

Theorem 15 is an immediate corollary of Proposition 5 and Proposition 6.

□

Proof of Theorems 3, 4, 5, 6 and 7

Theorems 3 and 4 can be syntactically proved using $\mathcal{AL}$ logic axiomatization. Theorem 5 proof is based on Theorem 3. As every $\mathcal{A}_{C:x}$ operator is normal, Theorems 6 and 7 can be proved by iteration of the Axiom (**K**) and the Rule of Necessitation (**Nec**) for every group of $2^{AGT\star}$ and every institution of $INST$. We provide in the sequel only the complete proof for Theorems (3a) and (3e).

Proof. Theorem (3a):

(1) $\vdash_{\mathcal{AL}} \mathcal{A}_{C:x}\bot \vee \neg\mathcal{A}_{C:x}\bot$, by (**ProTau**)

(2) $\vdash_{\mathcal{AL}} \neg\mathcal{A}_{C:x}\bot \rightarrow \mathcal{A}_{C:x}\neg\mathcal{A}_{C:x}\bot$, by (**NAccess**)

(3) $\vdash_{\mathcal{AL}} \mathcal{A}_{C:x}\bot \rightarrow \mathcal{A}_{C:x}\neg\mathcal{A}_{C:x}\bot$, by (**ProTau**), (**Nec**) and (**K**)

(4) $\vdash_{\mathcal{AL}} \mathcal{A}_{C:x}\neg\mathcal{A}_{C:x}\bot$, from (1), (2) and (3) by (**ProTau**)

□

Proof. Theorem (3e):

(1) $\vdash_{\mathcal{AL}} \mathcal{A}_{C:x}((\neg\mathcal{A}_{C:x}\bot \wedge \mathcal{A}_{C:x}\varphi) \rightarrow \mathcal{A}_{i:x}\varphi)$, for every $i \in C$, from Axiom (**Inc**) by inference rule (**Nec**),

(2) $\vdash_{\mathcal{AL}} \mathcal{A}_{C:x}((\neg\mathcal{A}_{C:x}\bot \wedge \mathcal{A}_{C:x}\varphi) \rightarrow \bigwedge_{i\in C} \mathcal{A}_{i:x}\varphi)$, from (1), by K principles

(3) $\vdash_{\mathcal{AL}} \mathcal{A}_{C:x}(\bigwedge_{i\in C} \mathcal{A}_{i:x}\varphi \rightarrow \varphi)$, from (**Unanim**)

(4) $\vdash_{\mathcal{AL}} \mathcal{A}_{C:x}((\neg\mathcal{A}_{C:x}\bot \wedge \mathcal{A}_{C:x}\varphi) \rightarrow \varphi)$, from (2) and (3) by (**ProTau**) and (**K**)

(5) $\vdash_{\mathcal{AL}} \mathcal{A}_{C:x}\neg\mathcal{A}_{C:x}\bot \rightarrow \mathcal{A}_{C:x}(\mathcal{A}_{C:x}\varphi \rightarrow \varphi)$, from (4) by (**ProTau**) and (**K**)

(6) $\vdash_{\mathcal{AL}} \neg\mathcal{A}_{C:x}\bot \rightarrow \mathcal{A}_{C:x}\neg\mathcal{A}_{C:x}\bot$, by (**NAccess**)

(7) $\vdash_{\mathcal{AL}} \neg\mathcal{A}_{C:x}\bot \rightarrow \mathcal{A}_{C:x}(\mathcal{A}_{C:x}\varphi \rightarrow \varphi)$, from (5) and (6) by (**ProTau**)

(8) $\vdash_{\mathcal{AL}} \mathcal{A}_{C:x}\bot \rightarrow \mathcal{A}_{C:x}(\mathcal{A}_{C:x}\varphi \rightarrow \varphi)$, by (**ProTau**), (**Nec**) and (**K**)

(9) $\vdash_{\mathcal{AL}} \mathcal{A}_{C:x}(\mathcal{A}_{C:x}\varphi \rightarrow \varphi)$, from (7) and (8) by (**ProTau**)

□

Proof of Proposition 1

Proof. Let suppose that the majority Principle (**Majority**) holds for any sets of agents C and B such that $B \subseteq C$ and $|C \setminus B| < |B|$. We will prove by induction on the set C_n, that there exists a set C_n such that:

$$(P_n) \qquad (\neg\mathcal{A}_{AGT:x}\bot \wedge \mathcal{A}_{AGT:x}\mathcal{A}_{\{i,j\}:x}\varphi) \rightarrow \mathcal{A}_{C_n:x}\varphi$$

where $i, j \in AGT$, $C_n \subseteq AGT$, $|C_n| = n$ and $n \geq 2$.

We begin by showing that (P_2) holds.

(1) $\vdash_{\mathcal{AL}} \neg\mathcal{A}_{AGT:x}\bot \wedge \mathcal{A}_{AGT:x}\mathcal{A}_{\{i,j\}:x}\varphi \rightarrow \mathcal{A}_{\{i,j\}:x}\mathcal{A}_{\{i,j\}:x}\varphi$, by (**Inc**).

(2) $\vdash_{\mathcal{AL}} \neg\mathcal{A}_{AGT:x}\bot \wedge \mathcal{A}_{AGT:x}\mathcal{A}_{\{i,j\}:x}\varphi \rightarrow \mathcal{A}_{\{i,j\}:x}\varphi$, from (1) by Theorem (3c)

(2) entails that (P_2) holds.

We suppose that (P_n) holds for any n such that $C_n \subset AGT$. Under this hypothesis we will show that (P_{n+1}) holds. We suppose that C_{n+1} is defined as: $C_{n+1} = C_n \cup \{i\}$, with $i \in AGT$ and $i \notin C_n$ (thus $C_n \subset C_{n+1}$).

(3) $\vdash_{\mathcal{AL}} (\neg\mathcal{A}_{AGT:x}\bot \wedge \mathcal{A}_{AGT:x}\mathcal{A}_{\{i,j\}:x}\varphi) \rightarrow \mathcal{A}_{C_n:x}\varphi$, by induction hypothesis (P_n)

(4) $\vdash_{\mathcal{AL}} (\mathcal{A}_{C_{n+1}:x}\neg\mathcal{A}_{AGT:x}\bot \wedge \mathcal{A}_{C_{n+1}:x}\mathcal{A}_{AGT:x}\mathcal{A}_{\{i,j\}:x}\varphi) \rightarrow \mathcal{A}_{C_{n+1}:x}\mathcal{A}_{C_n:x}\varphi$, from (3) by (**Nec**), (**K**) and standard properties of normal modal operator $\mathcal{A}_{C_{n+1}:x}$

(5) $\vdash_{\mathcal{AL}} (\neg\mathcal{A}_{AGT:x}\bot \wedge \mathcal{A}_{AGT:x}\mathcal{A}_{\{i,j\}:x}\varphi) \rightarrow \mathcal{A}_{C_{n+1}:x}\mathcal{A}_{C_n:x}\varphi \wedge \neg\mathcal{A}_{C_{n+1}:x}\bot$, from (4) by (**PAccess**), (**NAccess**), (**Mon**) and (**ProTau**)

(6) $\vdash_{\mathcal{AL}} \neg\mathcal{A}_{C_{n+1}:x}\bot \rightarrow \neg\mathcal{A}_{C_n:x}\bot$, by (**Mon**)

(7) $\vdash_{\mathcal{AL}} \mathcal{A}_{C_{n+1}:x}\neg\mathcal{A}_{C_{n+1}:x}\bot \rightarrow \mathcal{A}_{C_{n+1}:x}\neg\mathcal{A}_{C_n:x}\bot$, from (6) by (**Nec**), (**K**)

(8) $\vdash_{\mathcal{AL}} \neg\mathcal{A}_{C_{n+1}:x}\bot \rightarrow \mathcal{A}_{C_{n+1}:x}\neg\mathcal{A}_{C_n:x}\bot$, from (7) by (**NAccess**) and (**ProTau**)

(9) $\vdash_{\mathcal{AL}} (\neg\mathcal{A}_{AGT:x}\bot \wedge \mathcal{A}_{AGT:x}\mathcal{A}_{\{i,j\}:x}\varphi) \rightarrow \mathcal{A}_{C_{n+1}:x}(\mathcal{A}_{C_n:x}\varphi \wedge \neg\mathcal{A}_{C_n:x}\bot)$, from (5) and (8) by (**ProTau**) and standard properties of normal modal operator $\mathcal{A}_{C_{n+1}:x}$

(10) $\vdash_{\mathcal{AL}} (\neg\mathcal{A}_{AGT:x}\bot \wedge \mathcal{A}_{AGT:x}\mathcal{A}_{\{i,j\}:x}\varphi) \rightarrow \mathcal{A}_{C_{n+1}:x}(\bigwedge_{k\in C_n} \mathcal{A}_{k:x}\varphi)$, from (9) by (**Inc**), (**K**), (**Nec**) and (**ProTau**)

(11) $\vdash_{\mathcal{AL}} (\neg\mathcal{A}_{AGT:x}\bot \wedge \mathcal{A}_{AGT:x}\mathcal{A}_{\{i,j\}:x}\varphi) \rightarrow \mathcal{A}_{C_{n+1}:x}\varphi$, from (10) by (**Majority**), (**K**), (**Nec**) and (**ProTau**)

Thus (11) entails (P_{n+1}).

As (P_2) holds and from (P_n) we can infer that (P_{n+1}) for $n < |AGT|$, we can thus deduce by induction that (P_n) holds for $n \leq |AGT|$. In particular, we can deduce from the extension of the Principles (**Majority**) for every set of agents C and B such that $B \subseteq C$ and $|C \setminus B| < |B|$, that the following counterintuitive formula holds:

$$(\mathcal{A}_{AGT:x}\mathcal{A}_{\{i,j\}:x}\varphi \wedge \neg\mathcal{A}_{AGT:x}\bot) \rightarrow \mathcal{A}_{AGT:x}\varphi$$

□

Proof of Proposition 2

Lemma 7.

(7a) $\mathcal{A}_{C:x}\varphi \leftrightarrow Bel_i\mathcal{A}_{C:x}\varphi$ *if* $i \in C$

(7b) $\neg\mathcal{A}_{C:x}\varphi \leftrightarrow Bel_i\neg\mathcal{A}_{C:x}\varphi$ *if* $i \in C$

Proof. Lemma (7a) and (7b):

(1) $\neg\mathcal{A}_{C:x}\varphi \rightarrow Bel_i\neg\mathcal{A}_{C:x}\varphi$, by (**NegIntrAccept**), for $i \in C$

(2) $Bel_i\neg\mathcal{A}_{C:x}\varphi \rightarrow \neg Bel_i\mathcal{A}_{C:x}\varphi$, by Axiom (**D**) for Bel_i

(3) $Bel_i\mathcal{A}_{C:x}\varphi \rightarrow \mathcal{A}_{C:x}\varphi$, from (1), (2), and (**ProTau**), for $i \in C$

The proof of Lemma (7b) is similar to the one of Lemma (7a), we only use Axiom (**PIntrAccept**) instead of Axiom (**NegIntrAccept**). □

Proof. Propositions (2a) and (2b):

(1) $\mathcal{MB}_C\mathcal{A}_{C:x}\varphi \rightarrow \bigwedge_{i\in C} Bel_i(\mathcal{A}_{C:x}\varphi \wedge \mathcal{MB}_C\mathcal{A}_{C:x}\varphi)$, by (**FixPoint**)

(2) $\bigwedge_{i\in C} Bel_i(\mathcal{A}_{C:x}\varphi \wedge \mathcal{MB}_C\mathcal{A}_{C:x}\varphi) \rightarrow \bigwedge_{i\in C} Bel_i\mathcal{A}_{C:x}\varphi$, because Bel_i are normal modal operators

(3) $\bigwedge_{i\in C} Bel_i\mathcal{A}_{C:x}\varphi \rightarrow \mathcal{A}_{C:x}\varphi$, by Lemma (7a)

(4) $\mathcal{MB}_C\mathcal{A}_{C:x}\varphi \rightarrow \mathcal{A}_{C:x}\varphi$, from (1), (2), (3) by (**ProTau**)

(5) $\mathcal{A}_{C:x}\varphi \rightarrow Bel_i\mathcal{A}_{C:x}\varphi$, by (**PIntrAccept**), for every $i \in C$

(6) $\mathcal{A}_{C:x}\varphi \rightarrow E_C(\mathcal{A}_{C:x}\varphi \wedge \mathcal{A}_{C:x}\varphi)$, from (5), by (**ProTau**) and definition of E_C

(7) $\mathcal{A}_{C:x}\varphi \rightarrow \mathcal{MB}_C\mathcal{A}_{C:x}\varphi$, from (6) by inference rule (**InductionRule**) (left to right direction of Theorem (2a))

(8) $\mathcal{A}_{C:x}\varphi \leftrightarrow \mathcal{MB}_C\mathcal{A}_{C:x}\varphi$, from (4) and (7)

The proof of Proposition (2b) is similar to the one of Proposition (2a), we only use Lemma (7b) instead of Lemma (7a) and (**NegIntrAccept**) instead of (**PIntrAccept**). □

Proof of Theorem 8

To prove that these formulas are not valid in $\mathcal{AL}$, we only have to exhibit a model where there is a world where these formulas are false. We give the complete proof only for Theorem (8b), the others are very similar.

Proof. Theorem (8b):
We will build a $\mathcal{AL}$ model $\mathcal{M}$ in which there is a world w is which the formula is false, *i.e.*: $\mathcal{M}, w \vDash (\varphi_1 \overset{x}{\rhd} \varphi_2) \wedge (\varphi_2 \overset{x}{\rhd} \varphi_3) \wedge \neg(\varphi_1 \overset{x}{\rhd} \varphi_3)$. Let $ATM = \{\varphi_1, \varphi_2, \varphi_3\}$, $AGT = \{i\}$, $INST = \{x, y, z\}$ and $W = \{w, w_x, w_y, w_z\}$. We build the valuation function $\mathscr{V}$: $\mathscr{V}(\varphi_1) = \{w_y\}$, $\mathscr{V}(\varphi_2) = \{w_z\}$ and $\mathscr{V}(\varphi_3) = \{w_y\}$, and the relation $\mathscr{A}$: $\mathscr{A}_{\{i\}:x}(w) = \{w_x\}$, $\mathscr{A}_{\{i\}:y}(w) = \{w_y\}$ and $\mathscr{A}_{\{i\}:z}(w) = \{w_z\}$.

As we want $\mathcal{M}$ to be a $\mathcal{AL}$ model, we ensure that it satisfies the constraints **S.1**-**S.5**.

- In order to satisfy (**S.1**) and (**S.2**) we impose: $\langle w_y, w_x\rangle \in \mathscr{A}_{\{i\}:x}$, $\langle w_z, w_x\rangle \in \mathscr{A}_{\{i\}:x}$, $\langle w_x, w_y\rangle \in \mathscr{A}_{\{i\}:y}$, $\langle w_z, w_y\rangle \in \mathscr{A}_{\{i\}:y}$, $\langle w_x, w_z\rangle \in \mathscr{A}_{\{i\}:z}$ and $\langle w_y, w_z\rangle \in \mathscr{A}_{\{i\}:z}$;
- as there is only one agent in our model, (**S.3**) and (**S.5**) are satisfied;
- in order to satisfy (**S.4**) we impose that: $\langle w_x, w_x\rangle \in \mathscr{A}_{\{i\}:x}$, $\langle w_y, w_y\rangle \in \mathscr{A}_{\{i\}:y}$ and $\langle w_z, w_z\rangle \in \mathscr{A}_{\{i\}:z}$;

In this model $\mathcal{M}$:

- $\mathcal{M}, w \vDash [x](\varphi_1 \rightarrow \varphi_2) \wedge [x](\varphi_2 \rightarrow \varphi_3)$
- $\mathcal{M}, w \vDash \neg[y](\varphi_1 \rightarrow \varphi_2)$ and thus $\mathcal{M}, w \vDash \neg[Univ](\varphi_1 \rightarrow \varphi_2)$
- $\mathcal{M}, w \vDash \neg[z](\varphi_2 \rightarrow \varphi_3)$ and thus $\mathcal{M}, w \vDash \neg[Univ](\varphi_2 \rightarrow \varphi_3)$
- $\mathcal{M}, w \vDash [x](\varphi_1 \rightarrow \varphi_3) \wedge [y](\varphi_1 \rightarrow \varphi_3) \wedge [z](\varphi_1 \rightarrow \varphi_3)$, *i.e.* $\mathcal{M}, w \vDash [Univ](\varphi_1 \rightarrow \varphi_3)$

We have built a $\mathcal{AL}$ model which satisfies the formula $(\varphi_1 \overset{x}{\rhd} \varphi_2) \wedge (\varphi_2 \overset{x}{\rhd} \varphi_3) \wedge \neg(\varphi_1 \overset{x}{\rhd} \varphi_3)$. Thus, $(\varphi_1 \overset{x}{\rhd} \varphi_2) \wedge (\varphi_2 \overset{x}{\rhd} \varphi_3) \rightarrow (\varphi_1 \overset{x}{\rhd} \varphi_3)$ is not valid in $\mathcal{AL}$. By Theorem 1, we conclude that $(\varphi_1 \overset{x}{\rhd} \varphi_2) \wedge (\varphi_2 \overset{x}{\rhd} \varphi_3) \rightarrow (\varphi_1 \overset{x}{\rhd} \varphi_3)$ is not a theorem of $\mathcal{AL}$. □

Proof of Theorem 9

Proof. Theorems (9a) and (9b):
Since $[x]$ and $[Univ]$ are normal modal operators, they satisfy the rule of equivalence RE [Chellas, 1980]. Theorems (9a) and (9b) follow straightforwardly from RE. □

Proof. Theorem (9c):

(1) $\vdash_{\mathcal{AL}} ((\varphi_1 \to \varphi_2) \wedge (\varphi_1 \to \varphi_3)) \leftrightarrow (\varphi_1 \to (\varphi_2 \wedge \varphi_3))$, by (**ProTau**)

(2) $\vdash_{\mathcal{AL}} ([x](\varphi_1 \to \varphi_2) \wedge [x](\varphi_1 \to \varphi_3)) \to [x](\varphi_1 \to (\varphi_2 \wedge \varphi_3))$, from (1) by Theorem (6b)

(3) $\vdash_{\mathcal{AL}} (\neg[Univ](\varphi_1 \to \varphi_2) \wedge \neg[Univ](\varphi_1 \to \varphi_3)) \to (\neg[Univ](\varphi_1 \to \varphi_2) \vee \neg[Univ](\varphi_1 \to \varphi_3))$, by (**ProTau**)

(4) $\vdash_{\mathcal{AL}} (\neg[Univ](\varphi_1 \to \varphi_2) \vee \neg[Univ](\varphi_1 \to \varphi_3)) \to \neg[Univ]((\varphi_1 \to \varphi_2) \wedge (\varphi_1 \to \varphi_3))$, by standard properties of normal modal operator $[Univ]$

(5) $\vdash_{\mathcal{AL}} (\neg[Univ](\varphi_1 \to \varphi_2) \wedge \neg[Univ](\varphi_1 \to \varphi_3)) \to \neg[Univ](\varphi_1 \to (\varphi_2 \wedge \varphi_3))$, from (3) and (4) by (**ProTau**)

(6) $((\varphi_1 \overset{x}{\rhd} \varphi_2) \wedge (\varphi_1 \overset{x}{\rhd} \varphi_3)) \to (\varphi_1 \overset{x}{\rhd} (\varphi_2 \wedge \varphi_3))$, from (2) and (5) and (**ProTau**)

□

Proof. Theorems (9d) and (9e):
The proofs of Theorems (9d) and (9e) are very similar to the previous one. Both apply (**Nec**), (**K**) and propositional tautologies.

□

Proof of Theorem 10

All these theorems follow from the necessitation rule (**Nec**) and logical tautologies. Proofs also need Axiom (**K**) and theorems (M) and (C)[10] [Chellas, 1980] for the distribution over conjunction. We give the complete proof of Theorem (10b) as an example.

Proof. Theorem (10b):

(1) $\vdash_{\mathcal{AL}} ((\varphi_1 \to \varphi_2) \wedge (\varphi_2 \to \varphi_3)) \to (\varphi_1 \to \varphi_3)$, by (**ProTau**)

(2) $\vdash_{\mathcal{AL}} [x](((\varphi_1 \to \varphi_2) \wedge (\varphi_2 \to \varphi_3)) \to (\varphi_1 \to \varphi_3))$, from (1) by (**Nec**)

(3) $\vdash_{\mathcal{AL}} [x]((\varphi_1 \to \varphi_2) \wedge (\varphi_2 \to \varphi_3)) \to [x](\varphi_1 \to \varphi_3)$, from (2) by (**K**) and (C).

□

[10] The conjunction of both (M) and (C) give the equivalence: $[x](\varphi_1 \wedge \varphi_2) \leftrightarrow ([x]\varphi_1 \wedge [x]\varphi_2)$.

Proof of Theorem 11

Proof. Theorem (11a):

This theorem comes straightforwardly from Theorems 6a, 6b, 7a, 7b, and the following propositional tautologies:

(1) $\vdash_{\mathcal{AL}} (((\varphi \wedge \neg\psi) \rightarrow viol) \wedge (\neg\varphi \rightarrow viol)) \rightarrow (\neg\psi \rightarrow viol)$, by (**ProTau**)

(2) $\vdash_{\mathcal{AL}} (\neg\psi \rightarrow viol) \rightarrow ((\neg\psi \wedge \varphi) \rightarrow viol)$, by (**ProTau**)

□

Proof. Theorem (11b):

(1) $\vdash_{\mathcal{AL}} O_x\top \rightarrow \neg\,[Univ]\,(\bot \rightarrow viol)$, by the definition of $O_x\top$

(2) $\vdash_{\mathcal{AL}} \neg\,[Univ]\,(\bot \rightarrow viol) \rightarrow \neg\,[Univ]\,\top$, by (**ProTau**)

(3) $\vdash_{\mathcal{AL}} \neg\,[Univ]\,\top \rightarrow \bot$, by standard properties of normal modal operator $[Univ]$

(4) $\vdash_{\mathcal{AL}} \top \rightarrow \neg O_x\top$, from (1), (2) and (3) by (**ProTau**)

□

A logical account of institutions: from acceptances to norms via legislators

Emiliano Lorini and **Dominique Longin**
Institut de Recherche en Informatique de Toulouse (IRIT)
118 route de Narbonne
F-31062 Toulouse Cedex 04 (France)

Abstract

The aim of this paper is to provide a logical framework which enables reasoning about institutions and their dynamics. In our approach an institution is grounded on the *acceptances* of its members. We devote special emphasis to the role of *legislator*. We characterize the legislator as the role whose function is the creation and the modification of legal facts (e.g. permissions, obligations, *etc.*): the acceptance of the legislators that a certain norm is valid ensures that the norm is valid. The second part of the paper is devoted to the logical characterization of two important notions in the domain of legal and social theory: the notion of *constitutive rule* and the notion of *norm of competence*. A constitutive rule is a rule which is responsible for the creation of new kinds of (institutional) facts. A norm of competence is a rule which assigns powers to the agents playing certain roles within the institution. We show that norms of competence provide the criteria for institutional change.

Introduction

The problem of devising artificial institutions and modeling their dynamics is a fundamental problem in the multi-agent system domain (Dignum and Dignum 2001). Following (North 1990, p. 3), artificial institutions can be conceived as human-like: "the rules of the game in a society or the humanly devised constraints that structure agents' interaction". Starting from this concept of institution, many researchers working in the field of normative multi-agent systems have been interested in developing models which describe the different kinds of rules and norms that agents have to deal with. In some models of artificial institutions norms are conceived as means to achieve coordination among agents and agents are supposed to comply with them and to obey the authorities of the system as an end (Esteva, Padget, and Sierra 2001). More sophisticated models of institutions leave to the agents' autonomy the decision whether to comply or not with the specified rules and norms of the institution (Ågotnes et al. 2007; Lopez y Lopez, Luck, and d'Inverno 2004). However, all previous models abstract way from the legislative source of the norms of an institution, and from how institutions are created, maintained and changed by their members and not imposed from the outside by an external designer.

The aim of this work is to advance the state of the art on artificial institutions by proposing a logical model in which the existence and the dynamics of an institution (norms, rules, institutional facts, *etc.*) depend on the individual and collective attitudes of the agents which identify themselves as members of the institution. In particular, we propose a model in which an institution is grounded on the (individual and collective) *acceptances* of its followers and members, and its dynamics depend on the dynamics of these acceptances. On this aspect we agree with (Mantzavinos, North, and Shariq 2004), when the authors say that (p. 77):

> "only because institutions are anchored in peoples minds do they ever become behaviorally relevant. The *elucidation of the internal aspect is the crucial step* in adequately explaining the emergence, evolution, and effects of institutions." [Emphasis added].

In our model the agents are supposed to play certain social roles in one or more institutions and to accept things while playing these roles in the institutions. We devote special emphasis to the social role of *legislators* and show that the acceptances of the legislators directly affect the dynamics of the rules and the norms of the institution: the acceptances of the legislators are responsible for creating and modifying the obligations and the permissions of the institution.

It is worth noting that other authors in the MAS field have emphasized the need for a model which explains the origin and the evolution of institutions in terms of the agents' attitudes (Conte, Castelfranchi, and Dignum 1998; Conte and Dignum 2001; Boella and Van der Torre 2007). For instance, in agreement with (Hart 1992), Conte et al. (Conte, Castelfranchi, and Dignum 1998; Conte and Dignum 2001) have stressed that the existence of a norm in an institution (but also in a group, organization, *etc.*) depends on the recognition and acceptance of the norm by the members of the institution. In Conte et al.'s perspective, the agents contribute to the enforcement and the propagation of the norm. Furthermore, it has to be noted that, although in our approach institutions are anchored in agents' attitudes, we do not claim that institutions can be conceived as agents. Thus, our approach is different from (Boella and van der Torre 2004), in which the metaphor of normative systems as agents is used

and institutions are described in terms of mental attitudes such as beliefs and goals.

The paper is organized as follows. The second section of the article will be devoted to discussing the notion of acceptance and to distinguishing it from the classical notion of belief. In the third section we will introduce a modal logic which enables reasoning about obligations and individual and collective acceptances of agents while playing a certain *social role* within the institution (*i.e.* acceptance *qua* players of a certain role within the institution). This logic is an extension of the logic of acceptance we have presented in (Gaudou et al. 2008), in which social roles and obligations were not considered. We will devote special emphasis to the logical characterization of the acceptances of the agents playing the social role of *legislators* within the institution. We will formally characterize the legislators' power to create and modify the legal level of an institution. In the last sections of the article we will extend our analysis to the distinction between regulative and non-regulative components of an institution (Searle 1995). First, we will formalize the concept of *constitutive rule*, that is, the kind of rules accepted by the legislators which are responsible for the creation of new kinds of (institutional) facts. Since (Searle 1995; 1969) and (Jones and Sergot 1996), these rules have been expressed in terms of assertions of the form "X counts as Y in the context of institution x" (*e.g.* in the institutional context of US, a piece of paper with a certain shape, color, *etc.* counts as a five-dollar bill). We will conclude with a logical analysis of a particular form of constitutive rule, the so-called *norm of competence*. Norms of competence are rules which assign powers to the agents playing certain roles within the institution. We will show that norms of competence provide the criteria for institutional change.

An overview of the notion of *acceptance*

Before presenting our logical framework, we provide a brief overview of the concept of acceptance.

Whereas beliefs have been studied for decades, acceptances have only been examined since (Stalnaker 1984) and (Cohen 1992) while studying the nature of argument premises or reformulating Moore's paradox (Cohen 1992). If a belief that p is an attitude constitutively aimed at the truth of p, an acceptance is the output of "a decision to treat p as true in one's utterances and actions" (Hakli 2006; Bratman 1992) without being necessarily connected to the actual truth of the proposition. In order to better distinguish these two notions, it has been suggested (Hakli 2006) that while beliefs are not subject to the agent's will, acceptances are voluntary; while beliefs aim at truth, acceptances are sensitive to pragmatic considerations; while beliefs are shaped by evidence, acceptances need not be; while beliefs come in degrees, acceptances are qualitative; finally, while beliefs are context-independent, acceptance depends on context.

For the aims of this article we are particularly interested in the last feature, namely the fact that acceptances are context-dependent. In fact, one can decide (say for prudential reasons) to reason and act by "accepting" the truth of a proposition in a specific context, and possibly rejecting the very same proposition in another context. This aspect of acceptance has been studied both with respect to cooperative contexts (Gilbert 1989) (*e.g.* the context of a team) and with respect to institutional contexts (Tuomela 2007). We here continue the work initiated in (Gaudou et al. 2008) by exploring the role of acceptance with respect to institutional contexts. Institutional contexts are either rule-governed social practices (informal institutions) (*e.g.* language, games) or legal institutions, in which agents play certain social roles and on the background of which they reason. Consider a legal institution such as a trading company. The institutional context is the set of rules and norms which the agents conform to when they play the role of employees in the company. On the background of such contexts, we are interested in the individual and collective acceptances that can be formally captured. In the context of the trading company, for instance, the agents accept that something is true *qua* employees of the company. The state of acceptance *qua* player of a certain role in a certain institution is the kind of acceptance one is committed to when one is functioning as a player of a certain role in the institution (Tuomela 2007).

A logic of acceptance and obligation

Syntax

The syntactic primitives of our logic $\mathcal{L}$ of acceptance and obligation are the following: a finite set of $n > 0$ agents $AGT = \{i, j, \ldots\}$; a nonempty finite set of *atomic actions* $ACT = \{\alpha, \beta, \ldots\}$; a finite set of atomic formulas $ATM = \{p, q, \ldots\}$; a finite set of labels denoting institutional contexts $INST = \{x, y, \ldots\}$; a finite set of labels denoting social roles $ROLE = \{a, b, \ldots\}$. We suppose that $ROLE$ contains a (single) special role *leg* corresponding to the role of legislator of a certain institution. Moreover, we note $2^{AGT\star} = 2^{AGT} \setminus \{\emptyset\}$ the set of all nonempty subsets of agents, $\Delta = 2^{AGT\star} \times ROLE \times INST$ the set of all triples of non empty subsets of agents, social roles, and institutional contexts. We note $C{:}a{:}x$ the elements of Δ.

The language $\mathcal{LANG}$ of the logic $\mathcal{L}$ is defined by the following BNF:

$$\varphi ::= p \mid \neg\varphi \mid \varphi \vee \varphi \mid After_{i:\alpha}\varphi \mid [C{:}a{:}x]\,\varphi \mid O_x\varphi$$

where i ranges over AGT, α ranges over ACT, C ranges over $2^{AGT\star}$, a ranges over $ROLE$ and x ranges over $INST$. The classical boolean connectives $\wedge$, $\rightarrow$, $\leftrightarrow$, $\top$ and $\bot$ are defined from $\vee$ and $\neg$ in the usual manner. For notational convenience we write $[i{:}a{:}x]$ instead of $[\{i\}{:}a{:}x]$, for any $i \in AGT$.

Formula $[C{:}a{:}x]\,\varphi$ reads "the agents in group C accept that φ while playing role a together in the institution x". Operators of the form $[C{:}a{:}x]$ are extensions of the operators of acceptance $[C{:}x]$ we have introduced in (Gaudou et al. 2008) where we completely ignored social roles.

EXAMPLE 1. *$[C{:}activist{:}Greenpeace]\,protectEarth$ is read "the agents in C accept that their mission is to protect the Earth while playing together the role of activists in Greenpeace".*

The formula $[C{:}a{:}x]\,\bot$ has to be read "agents in C do not play role a together in the institution x" because we

assume that playing a role together in a certain institution is, at least in this minimal sense, a rational activity; conversely, $\neg\,[C{:}a{:}x]\,\bot$ has to be read "agents in C play role a together in the institution x"; $\neg\,[C{:}a{:}x]\,\bot \wedge [C{:}a{:}x]\,\varphi$ stands for "agents in C play role a together in institution x and they accept that φ while playing role a together in institution x" or simply "agents in C accept that φ *qua* players of role a in the institution x" (*i.e.* **collective acceptance**).

For the individual case, formula $\neg\,[i{:}a{:}x]\,\bot \wedge [i{:}a{:}x]\,\varphi$ has to be read "agent i accepts that φ *qua* player of role a in the institution x" (*i.e.* **individual acceptance**).

O_x are operators of obligation of standard deontic logic (SDL) indexed by institutional contexts and are used to express those facts which are legal with respect to a certain institution. Formula $O_x\varphi$ has to be read "φ is obligatory in the institution x". The dynamic operators of the form $After_{i:\alpha}$ are similar to the standard operators of dynamic logic (Harel, Kozen, and Tiuryn 2000) where both the action and its author are specified. Formula $After_{i:\alpha}\varphi$ has to be read "after agent i does action α, it is the case that φ".

We introduce four concepts by means of abbreviations. Their meanings will become clearer later in the analysis where the axioms and some theorems of the logic $\mathcal{L}$ will be discussed. For any $i \in AGT$, $\alpha \in ACT$, $C \in 2^{AGT\star}$ and $x \in INST$:

$$Happens_{i:\alpha}\varphi \stackrel{def}{=} \neg After_{i:\alpha}\neg\varphi$$

$$Leg(C{,}x) \stackrel{def}{=} \neg\,[C{:}leg{:}x]\,\bot \wedge \bigwedge_{C\subset B} [B{:}leg{:}x]\,\bot$$

$$Leg_x\varphi \stackrel{def}{=} \bigvee_{C\in 2^{AGT\star}} (Leg(C{,}x) \wedge [C{:}leg{:}x]\,\varphi)$$

$$Leg_{Univ}\varphi \stackrel{def}{=} \bigwedge_{x\in INST} Leg_x\varphi$$

Formula $Happens_{i:\alpha}\varphi$ has to be read "agent i performs action α and φ is true afterward". Formula $Leg(C{,}x)$ stands for "C is *the* group of legislators of institution x". Indeed, we suppose that the group of legislators of a certain institution x is the group C whose agents play together the role of legislators in x and there is no super-group B of C whose agents play together the role of legislators in x. We will show below that, for every institution x, there is only one group of legislators of x (this is the reason why we do not read $Leg(C{,}x)$ as "C is *a* group of legislators of institution x".) Formula $Leg_x\varphi$ stands for "there exists a group of legislators of x which accept φ". This can be shortened to "the legislators of x accept that φ" (due to the fact that in our logic every institution has only one group of legislators) or more simply "within the institutional context x, it is the case that φ". Finally, formula $Leg_{Univ}\varphi$ has to be read "the legislators of all institutions accept that φ" or simply "φ is universally accepted as true".

Semantics

We use a possible worlds semantics. A model of the logic $\mathcal{L}$ is a tuple $\mathcal{M} = \langle W, \mathscr{A}, \mathscr{R}, \mathscr{O}, \mathscr{V}\rangle$ where:

- W is a set of possible worlds;
- $\mathscr{A} : \Delta \longrightarrow (W \longrightarrow 2^W)$ associates each $C{:}a{:}x \in \Delta$ and world w with the set $\mathscr{A}_{C:a:x}(w)$ of worlds accepted by the group C at w, where the agents in C are playing role a in the institution x;
- $\mathscr{O} : INST \longrightarrow (W \longrightarrow 2^W)$ associates each $x \in INST$ and possible world w with the set $\mathscr{O}_x(w)$ of worlds which are ideal with regard to the institution x;
- $\mathscr{R} : AGT \times ACT \longrightarrow (W \longrightarrow 2^W)$ associates each agent $i \in AGT$, action $\alpha \in ACT$ and world w with the set $\mathscr{R}_{i:\alpha}(w)$ of worlds that are reachable from w through the occurrence of action α performed by i;
- $\mathscr{V} : W \longrightarrow 2^{ATM}$ is a truth assignment which associates each world w with the set $\mathscr{V}(w)$ of atomic propositions true in w.

The truth conditions of formulas are recursively defined as follows:

- $\mathcal{M}, w \models p$ iff $p \in \mathscr{V}(w)$;
- $\mathcal{M}, w \models \neg\varphi$ iff not $\mathcal{M}, w \models \varphi$;
- $\mathcal{M}, w \models \varphi \vee \psi$ iff $\mathcal{M}, w \models \varphi$ or $\mathcal{M}, w \models \psi$;
- $\mathcal{M}, w \models [C{:}a{:}x]\,\varphi$ iff for all $w' \in W$, if $w' \in \mathscr{A}_{C:a:x}(w)$ then $\mathcal{M}, w' \models \varphi$;
- $\mathcal{M}, w \models O_x\varphi$ iff for all $w' \in W$, if $w' \in \mathscr{O}_x(w)$ then $\mathcal{M}, w' \models \varphi$;
- $\mathcal{M}, w \models After_{i:\alpha}\varphi$ iff for all $w' \in W$, if $w' \in \mathscr{R}_{i:\alpha}(w)$ then $\mathcal{M}, w' \models \varphi$.

Axiomatization

Every operator of type $[C{:}a{:}x]$, O_x and $After_{i:\alpha}$ is supposed to be a normal modal operator satisfying standard axioms and rules of inference of the basic modal logic K. The rest of the section contains other axioms of the logic $\mathcal{L}$ and corresponding semantic constraints over $\mathcal{L}$ models.

Action. We suppose the following constraint over $\mathcal{L}$ models. For every $w \in W$, $i, j \in AGT$ and $\alpha, \beta \in ACT$:

if $w' \in \mathscr{R}_{i:\alpha}(w)$ and $w'' \in \mathscr{R}_{j:\beta}(w)$ then $w' = w''$ **S1**

The property **S1** says that all actions occurring in a world w lead to the same world. Thus, all actions occur in parallel and they do not have non-deterministic effects. This explains why we have phrased $Happens_{i:\alpha}\varphi$ "i does α and φ holds afterward" rather than "*it is possible that* i does α and φ holds afterward". Constraint **S1** corresponds to the following axiom of our logic. For every $i, j \in AGT$ and $\alpha, \beta \in ACT$:

$$Happens_{i:\alpha}\varphi \rightarrow After_{j:\beta}\varphi \qquad \textbf{Determ}$$

Acceptance and role playing. We suppose that: if agents in C accept that φ while playing role a together in the institution x then, for every subset B of C, while playing role a together in the institution x, the agents in B accept that the agents in C accept that φ, while playing role a together in the institution x. This means that given a group of agents C, all subgroups of C have access to all the facts that are

accepted by the agents in C while playing together a certain role in an institution. Such property is expressed by the following axiom. For every $B, C \in 2^{AGT\star}$, $x \in INST$ and $a \in ROLE$, if $B \subseteq C$ then:

$$[C{:}a{:}x]\,\varphi \rightarrow [B{:}a{:}x]\,[C{:}a{:}x]\,\varphi \qquad \mathbf{4_{Accept}}$$

Axiom $\mathbf{4_{Accept}}$ corresponds to the following semantic constraint over $\mathcal{L}$ models. For every $w \in W$, $B, C \in 2^{AGT\star}$, $x \in INST$ and $a \in ROLE$, if $B \subseteq C$ then:

$$\text{if } w' \in \mathscr{A}_{B:a:x}(w) \text{ then } \mathscr{A}_{C:a:x}(w') \subseteq \mathscr{A}_{C:a:x}(w) \qquad \mathbf{S2}$$

Moreover, we assume that if the agents in C accept that φ *qua* players of role a in the institution x then, for every subset B of C, it holds that the agents in B accept φ *qua* players of role a in the institution x. Thus, for every $B, C \in 2^{AGT\star}$, $x \in INST$ and $a \in ROLE$, if $B \subseteq C$ then:

$$(\neg\,[C{:}a{:}x]\bot \wedge [C{:}a{:}x]\,\varphi) \rightarrow (\neg\,[B{:}a{:}x]\,\bot \wedge [B{:}a{:}x]\,\varphi) \qquad \mathbf{Inc_{Accept}}$$

EXAMPLE 2. *Imagine three agents i, j, k that, qua Clue players, accept that someone called Mrs. Red, has been killed:* $\neg\,[\{i,j,k\}{:}player{:}Clue]\bot \wedge [\{i,j,k\}{:}player{:}Clue]\,killedMrsRed$. *This implies that also the two agents i, j qua Clue players accept that someone called Mrs. Red has been killed:* $\neg\,[\{i,j\}{:}player{:}Clue]\bot \wedge [\{i,j\}{:}player{:}Clue]\,killedMrsRed$.

Axiom $\mathbf{Inc_{Accept}}$ corresponds to the following semantic constraint over $\mathcal{L}$ models. For every $w \in W$, $B, C \in 2^{AGT\star}$, $x \in INST$ and $a \in ROLE$, if $B \subseteq C$ then:

$$\text{if } \mathscr{A}_{C:a:x}(w) \neq \emptyset \text{ then } \mathscr{A}_{B:a:x}(w) \neq \emptyset \text{ and } \mathscr{A}_{B:a:x}(w) \subseteq \mathscr{A}_{C:a:x}(w) \qquad \mathbf{S3}$$

The last axiom concerning acceptance and role playing says that: the agents in a group C, while playing together role a in the institution x, accept that they play together role a in the institution x. Formally for every $C \in 2^{AGT\star}$, $x \in INST$ and $a \in ROLE$:

$$[C{:}a{:}x]\,\neg\,[C{:}a{:}x]\,\bot \qquad \mathbf{RolePlay}$$

Intuitively, Axiom $\mathbf{RolePlay}$ means that if the agents in a group C play together a certain role within a certain institution then, this fact is public for the group C.

EXAMPLE 3. *Suppose that, during a concert, the agents in C play together the role of musicians in the context of the Philharmonic Orchestra. Then, this is public for the agents in C. That is, while playing together the role of musicians in the Philharmonic Orchestra, the agents in C accept that they are playing together the role of musicians in the Philharmonic Orchestra:* $[C{:}musician{:}Orchestra]\,\neg\,[C{:}musician{:}Orchestra]\,\bot$.

Axiom $\mathbf{RolePlay}$ corresponds to the following semantic constraint over $\mathcal{L}$ models. For every $w \in W$, $C \in 2^{AGT\star}$, $x \in INST$ and $a \in ROLE$:

$$\forall w' \in \mathscr{A}_{C:a:x}(w),\ \mathscr{A}_{C:a:x}(w') \neq \emptyset \qquad \mathbf{S4}$$

Acceptance and action. We also suppose the axiom of no forgetting for acceptance. This axiom describes how operators of acceptance interact with dynamic operators of the form $After_{i:\alpha}$. For every $i \in AGT$, $\alpha \in ACT$, $C \in 2^{AGT\star}$, $x \in INST$ and $a \in ROLE$:

$$(\,[C{:}a{:}x]\,After_{i:\alpha}\varphi \wedge \neg\,[C{:}a{:}x]\,After_{i:\alpha}\bot) \rightarrow After_{i:\alpha}\,[C{:}a{:}x]\,\varphi \qquad \mathbf{NF}$$

A lot of researchers have studied similar principles for the interaction between belief and action or between knowledge and action. Among them we should mention (Fagin et al. 1995; Gerbrandy 1999; Scherl and Levesque 2003; Herzig, Lang, and Polacsek 2000). It has to be noted that axiom $\mathbf{NF}$ relies on an assumption of complete and correct perception information. It is supposed that an agent i's action α occurs if and only if every group of agents (viz. single agent) is informed of this fact and updates its acceptances accordingly. Hence all action occurrences are supposed to be public. The axiom corresponds to the following semantic constraint over $\mathcal{L}$ models. For every $w \in W$, $i \in AGT$, $\alpha \in ACT$, $C \in 2^{AGT\star}$, $x \in INST$ and $a \in ROLE$:[1]

$$\text{if } \mathscr{A}_{C:a:x} \circ \mathscr{R}_{i:\alpha}(w) \neq \emptyset \text{ then } \mathscr{R}_{i:\alpha} \circ \mathscr{A}_{C:a:x}(w) \subseteq \mathscr{A}_{C:a:x} \circ \mathscr{R}_{i:\alpha}(w) \qquad \mathbf{S5}$$

Legal level of institutions. As far as the legal level of institutions is concerned, we suppose the standard deontic logic principle: if φ is obligatory in the context of institution x then, $\neg\varphi$ is not obligatory in the context of the same institution. Formally, for every $x \in INST$:

$$\neg(O_x\varphi \wedge O_x\neg\varphi) \qquad \boldsymbol{D_{Obl}}$$

Axiom $\boldsymbol{D_{Obl}}$ corresponds to the following semantic constraints over $\mathcal{L}$ models. For every $w \in W$ and $x \in INST$:

$$\mathscr{O}_x(w) \neq \emptyset \qquad \mathbf{S6}$$

Legislator. We add two specific axioms for the social role *legislator*. According to the first axiom, we cannot have two groups of agents which play *separately* the role of legislators in the same institution.

More precisely, given two groups B and C, if the agents in B play together the role of legislators in the institution x and the agents in C play together the role of legislators in the institution x, then the agents in $B \cup C$ play together the role of legislators in the institution x.

Formally, for every $B, C \in 2^{AGT\star}$ and $x \in INST$:

$$(\neg\,[B{:}leg{:}x]\,\bot \wedge \neg\,[C{:}leg{:}x]\,\bot) \rightarrow \neg\,[B \cup C{:}leg{:}x]\,\bot \qquad \mathbf{LegSum}$$

Axiom $\mathbf{LegSum}$ corresponds to the following semantic constraint over $\mathcal{L}$ models. For every $w \in W$, $B, C \in 2^{AGT\star}$ and $x \in INST$:

$$\text{if } \mathscr{A}_{B:leg:x}(w) \neq \emptyset \text{ and } \mathscr{A}_{C:leg:x}(w) \neq \emptyset \text{ then } \mathscr{A}_{B\cup C:leg:x}(w) \neq \emptyset \qquad \mathbf{S7}$$

[1] We note $\circ$ the standard composition operator such that, given two arbitrary functions $\mathscr{T}^1$ and $\mathscr{T}^2$ over worlds in W, $\mathscr{T}^1 \circ \mathscr{T}^2(w) = \bigcup \{\mathscr{T}^2(v) \mid v \in \mathscr{T}^1(w)\}$.

The following axiom **LegPower** is intended to capture the peculiar power of the group of legislators to create legal facts. We assume that the legislators of x accept that φ is obligatory in the institution x if and only if φ is obligatory in this institution. Formally, for every $x \in INST$:

$$Leg_x O_x \varphi \leftrightarrow O_x \varphi \qquad \textbf{LegPower}$$

It is worth noting that axiom **LegPower** do not express the intermediate step between the acceptance of the legislators that a norm is valid and the instantiation of the norm in the institution. This step, which is left implicit in the present analysis, is based on the legislators' act of *proclaiming* that the norm is valid.[2] Moreover, in axiom **LegPower** the legislators' power to create obligations is just expressed by means of a material implication (the left to right direction of the axiom). A more adequate characterization of this concept would require to substitute the material implication with a conditional which better expresses the fact that the legislators are *responsible* for the creation of the obligation.

Axiom **LegPower** corresponds to the following two semantic constraints over $\mathcal{L}$ models. For every $w \in W$, $C \in 2^{AGT\star}$ and $x \in INST$:

$$\begin{array}{c}\text{if } \mathscr{A}_{C:leg:x}(w) \neq \emptyset \text{ and } \forall B \text{ such that } C \subset B,\\ \mathscr{A}_{B:leg:x}(w) = \emptyset, \text{ then } \mathscr{O}_x(w) \subseteq \mathscr{A}_{C:leg:x} \circ \mathscr{O}_x(w)\end{array} \qquad \textbf{S8}$$

For every $w \in W$ and $x \in INST$:

$$\begin{array}{c}\exists C \in 2^{AGT\star} \text{ such that } \mathscr{A}_{C:leg:x}(w) \neq \emptyset \text{ and}\\ \forall B \text{ such that } C \subset B, \mathscr{A}_{B:leg:x}(w) = \emptyset \text{ and}\\ \mathscr{A}_{C:leg:x} \circ \mathscr{O}_x(w) \subseteq \mathscr{O}_x(w)\end{array} \qquad \textbf{S9}$$

We call $\mathcal{L}$ the logic axiomatized by the principles presented above and we write $\vdash_{\mathcal{L}} \varphi$ iff formula φ is a theorem of $\mathcal{L}$ provable from our axioms by the inference rules of modus ponens and necessitation for every modal operator. Moreover, we write $\models_{\mathcal{L}} \varphi$ iff formula φ is *valid* in all $\mathcal{L}$ models, *i.e.* $\mathcal{M}, w \models \varphi$ for every $\mathcal{L}$ model $\mathcal{M}$ and world w in $\mathcal{M}$. Finally, we say that a formula φ is *satisfiable* if $\not\models_{\mathcal{L}} \neg\varphi$ We can prove that the logic $\mathcal{L}$ is *sound* and *complete* with respect to the class of $\mathcal{L}$ models. Namely:

Theorem 1 *$\mathcal{L}$ is determined by the class of $\mathcal{L}$ models.*

Properties of acceptances

This section provides further clarifications of the concept of acceptance. In particular, we focus on the distinction between acceptance and belief.

As said, there is a large literature about the distinction between belief and acceptance. For us, belief and acceptance are clearly different concepts in several senses. (For convenience, we adopt $Bel_i\varphi$ as a notation for "the agent i believes that φ is true", and we suppose that these operators are defined as usual in a KD45 modal logic).

Individual belief and individual acceptance are both private mental attitudes but: a belief does not depend on contexts, whilst an acceptance is a context-dependent attitude which is entertained by an agent *qua* player of a certain role within a given institution. Therefore, an agent can privately disbelieve something he accepts while playing a certain role within a given institution. Formally: $Bel_i\varphi \wedge [i{:}a{:}x]\neg\varphi$ should be satisfiable. In a similar way, as emphasized in (Tuomela 1992), a collective acceptance that φ of a group of agents C (*qua* players of a certain role within a given institution) might be compatible with the fact that every agent in C does not believe that φ (or that every agent in C believes that $\neg\varphi$). The following example, inspired from (Tuomela 1992, p. 285), illustrates this.

EXAMPLE 4. *At the end of the 80s, the Communist Party of Ruritania accepted that capitalist countries will soon perish (but none of its members really believed so).*

We can formalize the example as follows: $\neg\,[C{:}member{:}CPR]\,\bot \;\wedge\; [C{:}member{:}CPR]\,ccwp \;\wedge\; \bigwedge_{i\in C} \neg Bel_i\, ccwp$. This means that the agents in C accept that capitalist countries will perish ($ccwp$) *qua* members of the Communist Party of Ruritania (CPR) but nobody in C believes this.

Properties of legislators

This section is devoted to studying the notion of legislator. The following theorem highlights some of its properties.

Theorem 2 *For every $x \in INST$ and $B, C \in 2^{AGT\star}$ such that $B \neq C$:*

$$\vdash_{\mathcal{L}} Leg(C,x) \rightarrow \neg Leg(B,x) \qquad (2a)$$

$$\vdash_{\mathcal{L}} \bigvee_{C\in 2^{AGT\star}} Leg(C,x) \qquad (2b)$$

$$\textit{If } \vdash_{\mathcal{L}} \varphi \textit{ then } \vdash_{\mathcal{L}} Leg_x\varphi \qquad (2c)$$

$$\vdash_{\mathcal{L}} (Leg_x(\varphi \rightarrow \psi) \wedge Leg_x\varphi) \rightarrow Leg_x\psi \qquad (2d)$$

$$\vdash_{\mathcal{L}} \neg(Leg_x\varphi \wedge Leg_x\neg\varphi) \qquad (2e)$$

$$\vdash_{\mathcal{L}} Leg_x\varphi \rightarrow Leg_xLeg_x\varphi \qquad (2f)$$

Theorem 2a ensures that, for every institution, there is only one group of legislators. Theorem 2b says that for every institution x, there is always a group of legislators of x. Theorems 2c–2f highlight the fact that operators of type Leg_x are normal modal operators satisfying the axioms and rules of inference of the system KD4 (Chellas 1980). In particular, Theorem 2f says that, if the legislators of x accept that φ, then the legislators of x accept that the legislators of x accept that φ. This latter property captures a sort of 'introspective' capacity of legislators: legislators have access to those facts that they accept *qua* legislators.

Weak permissions *vs.* strong permissions. As discussed above, the legislator of a certain institution is the social role which has the function of creating and modifying legal facts. In particular, we have assigned to the legislators the power to create obligations (Axiom **LegPower**). Now, let us consider permissions in order to establish a new formal relationship between the legal level of institutions and the legislators. We define permission in the usual way by taking the dual of the operator of obligation. We say that "φ is something permitted in the institutional context x" (noted $P_x\varphi$) if

[2] For a logical characterization of the act of *proclaiming*, see (Gelati et al. 2004).

and only if $\neg\varphi$ is not obligatory in the institutional context x. Formally:

$$P_x\varphi \stackrel{def}{=} \neg O_x \neg\varphi.$$

The following theorem can be proved.

Theorem 3 *For every $x \in INST$:*

$$\vdash_{\mathcal{L}} Leg_x P_x\varphi \rightarrow P_x\varphi$$

Thus, in our logic the legislators are also endowed with the power of creating permissions. It has to be noted that the converse of Theorem 3 is not a theorem of our logic: $P_x\varphi \wedge \neg Leg_x P_x\varphi$ is satisfiable in the logic $\mathcal{L}$.

Thus, φ might be permitted within institution x while the legislators of x do not accept φ to be permitted within the context x. (In this sense, permissions behave differently from obligations, cf. Axiom **LegPower**.)

This property is justified by the distinction between weak permission and strong permission, which was emphasized by several authors in analytical philosophy (Alchourrón and Bulygin 1971; Raz 1975; Von Wright 1963) and in the domain of normative MAS (Boella and van der Torre 2003). According to Von Wright for instance "[...] An act will be said to be permitted in the weak sense if it is not forbidden; and it will be said to be permitted in the strong sense if it is not forbidden but subject to norm." (Von Wright 1963, p. 86). A weak permission corresponds to the absence in the institution of a norm prohibiting φ. This concept is captured by the formula $P_x\varphi$ of our logic. A strong permission corresponds to the existence in the institution of an explicit norm, accepted by the legislators, according to which φ is permitted, which is captured by the formula $Leg_x P_x\varphi$ of our logic. In this perspective, Theorem 3 states that a strong permission implies a weak permission. In contrast, the converse is not valid.

In the rest of the article we will investigate the fundamental concepts of *constitutive rule*, *norm of competence* and *institutionalized power* within the formal framework of $\mathcal{L}$.

From constitutive rules to norms of competence

According to many philosophers working on social theory (Rawls 1955; Alchourrón and Bulygin 1971) and researchers in the field of normative multi-agent systems (Boella and van der Torre 2004), institutions are based both on regulative as well as constitutive (*i.e.* non-regulative) components. That is, institutions are not only defined in terms of sets of permissions, obligations, and prohibitions (*i.e. norms of conduct*) but also in terms of rules which specify and create new forms of behavior and concepts. According to Searle for instance "[...] regulative rules regulate antecedently or independently existing forms of behavior [...]. But constitutive rules do not merely regulate, they create or define new forms of behavior" (Searle 1969, p. 33). In Searle's theory of institutions (Searle 1969; 1995), constitutive rules are expressed by means of "counts-as" assertions of the form "X counts as Y in context x" where the context x refers to the institution/normative system in which the rule is specified. For example, in the institutional context of US, a piece of paper with a certain shape, color, *etc.* counts as a five-dollar bill.

The distinction between regulative rules and constitutive rules can be expressed in our formal language $\mathcal{L}$. Regulative rules are characterized in $\mathcal{L}$ by the constructions $O_x\varphi$ (obligation), $P_x\varphi$ (weak permission) and $Leg_x P_x\varphi$ (strong permission) introduced above.

The following two subsections are devoted to presenting a formal characterization of the concept of constitutive rule. We will first provide a formal analysis of the general notion of constitutive rule. Then, we will investigate a particular form of constitutive rule which is commonly referred to as *norm of competence* (Bulygin 1992). A norm of competence of a certain institution x is a norm on the basis of which special (institutionalized) powers are assigned to the agents playing a certain role in the institution.

Constitutive rules

A notion of *constitutive rule* of the form "φ counts as ψ in the institutional context x" can be defined in our logic $\mathcal{L}$ by means of the operator Leg_x. We conceive a constitutive rule as a material implication of the form $\varphi \rightarrow \psi$ in the scope of an operator Leg_x. Thus, "φ counts as ψ in the institutional context x" only if the legislators of institution x accept that φ entails ψ. Furthermore, we suppose that a constitutive rule is intrinsically contextual, that is, a rule that is not universally valid while it is accepted by the legislators of a certain institution. More precisely, we exclude the situation in which $Leg_{Univ}(\varphi \rightarrow \psi)$ is true (the legislators of every institution accept that φ entails ψ). More generally, for every $x \in INST$ the following abbreviation $\varphi \stackrel{x}{\rhd} \psi$ (that stands for "φ counts as ψ in the institutional context x") is given:

$$\varphi \stackrel{x}{\rhd} \psi \stackrel{def}{=} Leg_x(\varphi \rightarrow \psi) \wedge \neg Leg_{Univ}(\varphi \rightarrow \psi)$$

EXAMPLE 5. *The formula $sixteen \stackrel{Brazil}{\rhd} votingAge$ stands for "in Brazil, the fact that a person is sixteen year old counts as the fact that he has the voting age". This means that the legislators of Brazil accept that being sixteen year old entails having the voting age,* i.e. $Leg_{Brazil}(sixteen \rightarrow votingAge)$, *and there are legislators of other countries who do not accept this,* i.e. $\neg Leg_{Univ}(sixteen \rightarrow votingAge)$. *Indeed, there are other countries such as Italy and France in which the voting age is set at eighteen years and not at sixteen.*[3]

It is interesting to note that $\varphi \stackrel{x}{\rhd} \psi$ satisfies some intuitive properties of counts-as conditionals as isolated in (Jones and Sergot 1996).

[3]Note that a more precise characterization of this example requires a quantification over the set of agents AGT. That is, the constitutive rule should be specified by the formula $\bigwedge_{i \in AGT}(sixteen(i) \stackrel{Brazil}{\rhd} votingAge(i))$ which is meant to stand for " in Brazil, for every agent i, i is sixteen year old counts as i has the voting age".

Theorem 4 *For every $x \in INST$:*

$$\textit{If } \vdash_{\mathcal{L}} (\varphi_2 \leftrightarrow \varphi_3) \textit{ then } \vdash_{\mathcal{L}} (\varphi_1 \overset{x}{\rhd} \varphi_2 \leftrightarrow \varphi_1 \overset{x}{\rhd} \varphi_3) \quad (4a)$$

$$\textit{If } \vdash_{\mathcal{L}} (\varphi_1 \leftrightarrow \varphi_3) \textit{ then } \vdash_{\mathcal{L}} (\varphi_1 \overset{x}{\rhd} \varphi_2 \leftrightarrow \varphi_3 \overset{x}{\rhd} \varphi_2) \quad (4b)$$

$$\vdash_{\mathcal{L}} (\varphi_1 \overset{x}{\rhd} \varphi_2 \wedge \varphi_1 \overset{x}{\rhd} \varphi_3) \rightarrow (\varphi_1 \overset{x}{\rhd} (\varphi_2 \wedge \varphi_3)) \quad (4c)$$

$$\vdash_{\mathcal{L}} (\varphi_1 \overset{x}{\rhd} \varphi_2 \wedge \varphi_3 \overset{x}{\rhd} \varphi_2) \rightarrow ((\varphi_1 \vee \varphi_3) \overset{x}{\rhd} \varphi_2) \quad (4d)$$

$$\vdash_{\mathcal{L}} (\varphi_1 \overset{x}{\rhd} \varphi_2 \wedge (\varphi_1 \wedge \varphi_2) \overset{x}{\rhd} \varphi_3) \rightarrow (\varphi_1 \overset{x}{\rhd} \varphi_3) \quad (4e)$$

For instance, Theorem 4e corresponds to a property of cumulative transitivity (cut). We can easily show that the operator $\overset{x}{\rhd}$ does not satisfy reflexivity, transitivity and weakening of the antecedent, that is: $\varphi \overset{x}{\rhd} \varphi$, $(\varphi_1 \overset{x}{\rhd} \varphi_2 \wedge \varphi_2 \overset{x}{\rhd} \varphi_3) \rightarrow \varphi_1 \overset{x}{\rhd} \varphi_3$, and $\varphi_1 \overset{x}{\rhd} \varphi_2 \rightarrow (\varphi_1 \wedge \varphi_3) \overset{x}{\rhd} \varphi_2$ are not valid in $\mathcal{L}$. This is due to the "local" nature of the operator $\overset{x}{\rhd}$. For instance, $\varphi_1 \overset{x}{\rhd} \varphi_2$ and $\varphi_2 \overset{x}{\rhd} \varphi_3$ might be constitutive rules of the institution x, while $\varphi_1 \overset{x}{\rhd} \varphi_3$ fails to be a constitutive rule of x since it is not intrinsically contextual (*i.e.* $Leg_{Univ}(\varphi_1 \overset{x}{\rhd} \varphi_3)$ holds).

It has to be noted that our notion of "counts as" is similar to the notion of *proper classificatory rule* defined in (Grossi, Meyer, and Dignum 2006). [4]

Norms of competence and institutionalized power

According to some legal theorists (Bulygin 1992; Searle 1969; Hart 1992), norms of competence are power-conferring rules which should not be reduced to norms of conduct such as obligations, prohibitions, commands and permissions. These kinds of rules assign special powers to the agents playing certain roles within the institution. They have a fundamental function in normative and legal systems since they provide the criteria for institutional change, that is, they provide the criteria for the creation and modification of institutional facts (*e.g.* agent i and agent j are married, this house is i's property, *etc.*) and normative facts (*e.g.* obligations and permissions).

Norms of competence can be specified in the logic $\mathcal{L}$ in the following way. For any $x \in INST$, $a \in ROLE$ and $\alpha \in ACT$:

$$Power(a,\alpha,\varphi,x) \overset{def}{=} \bigwedge_{i \in AGT} (\neg [i{:}a{:}x] \bot \overset{x}{\rhd} After_{i:\alpha}\varphi)$$

$Power(a,\alpha,\varphi,x)$ reads "in institution x there is a norm of competence which assigns to the agents playing role a in x the power to ensure φ by performing action α" or simply "in the institutional context x, the agents playing role a (in x) have the power to ensure φ by performing action α".

EXAMPLE 6. *The formula $Power(priest,gesture,married,church)$ is meant to stand for "in the institutional context of Catholic Church, the agents playing the role of priest (in the Church) have the power of marrying a couple by performing certain gestures".*

From the previous concept of *institutionalized power*, we can define a corresponding notion of *exercise of institutionalized power*. We say that in the institutional context x, an agent i playing role a (in x) *exercises* its power of ensuring φ by doing action α if and only if:

i. in context x, the agents playing role a (in x) have the power to ensure φ by performing action α;

ii. the legislators of x accept that i is playing role a in x and that agent i performs action α.

Formally, for any $x \in INST$, $a \in ROLE$, $i \in AGT$ and $\alpha \in ACT$:

$$ExPower(i,a,\alpha,\varphi,x) \overset{def}{=}$$
$$Power(a,\alpha,\varphi,x) \wedge Leg_x(\neg [i{:}a{:}x] \bot \wedge Happens_{i:\alpha} \top)$$

Our aim is to show how the exercise of an institutionalized power by an agent modifies the current structure of the institution through the creation of new institutional facts. To this end, we have to introduce the following definition. For any $x \in INST$, $i \in AGT$ and $\alpha \in ACT$:

$$NoChange(x, i{:}\alpha) \overset{def}{=}$$
$$\textstyle\bigwedge_{C \in 2^{AGT\star}} (Leg(C,x) \rightarrow After_{i:\alpha} Leg(C,x))$$

$NoChange(x, i{:}\alpha)$ is meant to stand for "the group of legislators of x do not change after agent i performs action α" (*i.e.* for any $C \in 2^{AGT\star}$, if C is the group of legislators of x then, after agent i performs action α, C is still the group of legislators of x).

We are now in the position to prove two theorems which highlight the dynamic aspect of institutions based on the exercise of institutionalized power.

Theorem 5 *For every $x \in INST$, $a \in ROLE$, $i \in AGT$ and $\alpha \in ACT$:*

$$\vdash_{\mathcal{L}} (ExPower(i,a,\alpha,\varphi,x) \wedge NoChange(x, i{:}\alpha)) \rightarrow After_{i:\alpha} Leg_x \varphi$$

According to Theorem 5, if in the institution x agent i playing role a exercises its power of ensuring φ by doing action α, under the condition that the legislators of x do not change after agent i performs action α then, after i performs α, it is the case that φ is true within the institution x.

EXAMPLE 7. *Suppose that in the Church, agent i playing the role of priest exercises its power of marrying a couple by performing certain gestures, noted $ExPower(i,priest,gesture,married,church)$. Then, under the condition $NoChange(church,i{:}gesture)$ (the group of legislators of the Church do not change after i's action), after i performs the gestures, the couple will be married in the context of the Church, noted $After_{i:gesture} Leg_{church} married$.*

Theorem 6 *For every $x \in INST$, $a \in ROLE$, $i \in AGT$ and $\alpha \in ACT$:*

$$\vdash_{\mathcal{L}} (NoChange(x, i{:}\alpha) \wedge ExPower(i,a,\alpha,O_x\varphi,x)) \rightarrow After_{i:\alpha} O_x \varphi$$

[4] We refer to (Grossi, Meyer, and Dignum 2006) for interesting arguments why proper classificatory rules should not necessarily satisfy reflexivity, transitivity and weakening of the antecedent.

Theorem 6 highlights the dynamics of obligations in an institution due to the exercise of institutionalized powers by agents. It says that: if in the institution x agent i playing role a exercises its power of creating the obligation that φ by doing action α, then, after i performs α, it is the case that φ is obligatory in x. (Under the condition that the legislators of x do not change.)

Conclusion

We have presented in this article a logic of acceptance and obligation and applied it to the analysis of institutions and their dynamics. Our logic of acceptance and obligation allows to express that certain agents accept something to be true *qua* players of a role within an institution. We have devoted special emphasis to the social role *legislator* by discussing its influence on the creation and the modification of the norms of an institution. In the second part of the paper we have formalized the concept of constitutive rule, that is, a rule of the form "X counts as Y in the context of institution x" which is responsible for the creation of institutional facts. While constitutive rules are usually defined from the external perspective of a normative system or institution, we have, once again, anchored these rules in the acceptances of the legislators. We have concluded with an analysis of a particular form of constitutive rule, the so-called norm of competence. A norm of competence is a norm on the basis of which special institutionalized powers are assigned to the agents playing certain roles in the institution.

Directions for future research are manifold. Our future works will be devoted to better clarify the relationships between the legislators of an institution and the acceptances of the members of the institution. In particular, we will integrate the following two general principles into our logical framework. According to first principle, all members of an institution have to accept, *qua* members of the institution, all facts which are accepted by the legislators of the institution. This principle expresses that the members of an institution are necessarily subject to what the legislators of the institution accept. According to the second principle, the agents in a set C are the legislators of a certain institution only if the members of the institution accept that the agents in C are the legislators of the institution and recognize them as the legislators. This second principle expresses that the legitimacy of the legislator's authority is necessarily based on the consent of the members of the institution. Our logic is sufficiently expressive to capture the two principles:

- $[C{:}a{:}x]\, Leg_x\varphi \rightarrow [C{:}a{:}x]\, \varphi$
- $Leg(B,x) \rightarrow [C{:}a{:}x]\, Leg(B,x)$

The first formula expresses that, if the agents in C, while playing role a in institution x, accept that the legislators of x accept φ then, the agents in C have to accept φ while playing role a in x. The second formula expresses that, if B is the group of legislators of institution x then, for every set of agents C and role a, the agents in C, while playing role a in x, have to accept that B is the group of legislators of x.

Furthermore, in future extensions of this work, we will investigate the decision to join or not to join (and the decision to leave or to remain member of) a given institution. This decision is influenced by the inconsistency between the agent's preferences and goals and the current norms and rules of the institution. For instance, if the agent's goals conflict with the norms proclaimed by the legislators then, the agent will probably decide not to join the institution. In order to model this form of reasoning, we will extend our logical framework to modalities expressing agents' goals and preferences, such as the ones provided in (Cohen and Levesque 1990).

Acknowledgments

Emiliano Lorini is financed by the project ForTrust (Social Trust Analysis and Formalization) funded by the french Agence Nationale de la Recherche (ANR). Thanks are due to our colleague Andreas Herzig and to the anonymous reviewers of this paper for their helpful comments and suggestions.

Appendix: proofs of theorems

Proof of Theorem 1. It is a routine to prove soundness, whereas completeness is obtained by Sahlqvist completeness theorem (Blackburn, de Rijke, and Venema 2001). Indeed, all axioms of the logic $\mathcal{L}$ are in the Sahlqvist class.

Proof of Theorem 2a.

$$\vdash_{\mathcal{L}} Leg(C,x) \rightarrow \neg Leg(B,x) \text{ if } B \neq C$$

It is enough to prove the Theorem for the following two cases.

- CASE 1. $B \subset C$ or $C \subset B$
- CASE 2. $B \not\subset C$ and $C \not\subset B$

The proof for the first case is straightforward. Let us give a proof for the second case.

1. $(Leg(C,x) \wedge Leg(B,x)) \rightarrow$
   $(\neg\, [C{:}leg{:}x] \bot \wedge \bigwedge_{C \subset D} [D{:}leg{:}x] \bot \wedge \neg\, [B{:}leg{:}x] \bot)$
   from def. $Leg(C,x)$ and def. $Leg(B,x)$
2. $(\neg\, [C{:}leg{:}x] \bot \wedge \bigwedge_{C \subset D} [D{:}leg{:}x] \bot \wedge \neg\, [B{:}leg{:}x] \bot) \rightarrow$
   $(\neg\, [B \cup C{:}leg{:}x] \bot \wedge \bigwedge_{C \subset D} [D{:}leg{:}x] \bot)$
   From Axiom **LegSum**
3. $(\neg\, [B \cup C{:}leg{:}x] \bot \wedge \bigwedge_{C \subset D} [D{:}leg{:}x] \bot) \rightarrow \bot$
   From the facts $B \neq C$, $B \neq \emptyset$, $C \neq \emptyset$[5], $B \not\subset C$ and $C \not\subset B$
4. $(Leg(C,x) \wedge Leg(B,x)) \rightarrow \bot$
   From 1,2,3
5. $Leg(C,x) \rightarrow \neg Leg(B,x)$ From 4

Proof of Theorem 2b.

$$\vdash_{\mathcal{L}} \bigvee_{C \in 2^{AGT^\star}} Leg(C,x)$$

1. $O_x \top$

[5] Remember that B and C are member of the set $2^{AGT^\star}$ of non empty subsets of agents.

2. $O_x\top \rightarrow Leg_x\top$
From Axiom $\mathbf{LegPower}$
3. $Leg_x\top$
From 1,2
4. $Leg_x\top \rightarrow (\bigvee_{C\in 2^{AGT\star}} Leg(C,x))$
From def. $Leg_x\top$ and def. $Leg(C,x)$
5. $\bigvee_{C\in 2^{AGT\star}} Leg(C,x)$
From 3,4

Proof of Theorem 2c.

$$\text{From } \vdash_{\mathcal{L}} \varphi \text{ infer } \vdash_{\mathcal{L}} Leg_x\varphi$$

Let us suppose that φ is a theorem of $\mathcal{L}$. We prove that $Leg_x\varphi$ is a theorem of $\mathcal{L}$ as well.

1. φ
From hypothesis
2. $\bigwedge_{C\in 2^{AGT\star}} [C{:}leg{:}x]\,\varphi$
From 1 and necessitation rule for $[C{:}leg{:}x]$
3. $\bigvee_{C\in 2^{AGT\star}} Leg(C,x)$
From Theorem 2b
4. $\bigvee_{C\in 2^{AGT\star}} (Leg(C,x) \wedge [C{:}leg{:}x]\,\varphi)$
From 2,3
5. $Leg_x\varphi$
From 4 and def. $Leg_x\varphi$

Proof of Theorem 2d.

$$\vdash_{\mathcal{L}} (Leg_x(\varphi \rightarrow \psi) \wedge Leg_x\varphi) \rightarrow Leg_x\psi$$

1. $Leg_x(\varphi \rightarrow \psi) \leftrightarrow$
$\bigvee_{C\in 2^{AGT\star}} (Leg(C,x) \wedge [C{:}leg{:}x]\,(\varphi \rightarrow \psi))$
From def. $Leg_x(\varphi \rightarrow \psi)$
2. $Leg_x\varphi \leftrightarrow \bigvee_{C\in 2^{AGT\star}} (Leg(C,x) \wedge [C{:}leg{:}x]\,\varphi)$
From def. $Leg_x\varphi$
3. $\bigwedge_{B,C\in 2^{AGT\star}, B\neq C} (Leg(C,x) \rightarrow \neg Leg(B,x))$
From Theorem 2a
4. $(Leg_x(\varphi \rightarrow \psi) \wedge Leg_x\varphi) \rightarrow \bigvee_{C\in 2^{AGT\star}} (Leg(C,x) \wedge$
$[C{:}leg{:}x]\,(\varphi \rightarrow \psi) \wedge [C{:}leg{:}x]\,\varphi)$
From 1,2,3
5. $\bigvee_{C\in 2^{AGT\star}} (Leg(C,x) \wedge [C{:}leg{:}x]\,(\varphi \rightarrow \psi) \wedge$
$[C{:}leg{:}x]\,\varphi) \rightarrow \bigvee_{C\in 2^{AGT\star}} (Leg(C,x) \wedge [C{:}leg{:}x]\,\psi)$
From Axiom $\mathbf{K}$ for $[C{:}leg{:}x]$
6. $\bigvee_{C\in 2^{AGT\star}} (Leg(C,x) \wedge [C{:}leg{:}x]\,\psi) \rightarrow Leg_x\psi$
From def. $Leg_x\psi$
7. $(Leg_x(\varphi \rightarrow \psi) \wedge Leg_x\varphi) \rightarrow Leg_x\psi$
From 4,5,6

Proof of Theorem 2e.

$$\vdash_{\mathcal{L}} \neg(Leg_x\varphi \wedge Leg_x\neg\varphi)$$

1. $\bot \rightarrow O_x\bot$
From standard principles of propositional calculus
2. $Leg_x\bot \rightarrow Leg_xO_x\bot$
From 1, Theorems 2c and 2d
3. $Leg_xO_x\bot \rightarrow O_x\bot$
From Axiom $\mathbf{LegPower}$
4. $O_x\bot \rightarrow (O_x\varphi \wedge O_x\neg\varphi)$
5. $(O_x\varphi \wedge O_x\neg\varphi) \rightarrow \bot$
From Axiom $\boldsymbol{D_{Obl}}$
6. $Leg_x\bot \rightarrow \bot$
From 2,3,4,5
7. $\neg Leg_x(\varphi \wedge \neg\varphi)$
From 6
8. $\neg(Leg_x\varphi \wedge Leg_x\neg\varphi)$
From 7 and standard principles of the normal modal operator Leg_x (*i.e.* $(Leg_x\varphi \wedge Leg_x\psi) \leftrightarrow Leg_x(\varphi \wedge \psi)$ is a theorem of $\mathcal{L}$)

Proof of Theorem 2f.

$$\vdash_{\mathcal{L}} Leg_x\varphi \rightarrow Leg_xLeg_x\varphi$$

1. $Leg_x\varphi \rightarrow \bigvee_{C\in 2^{AGT\star}} (\neg\,[C{:}leg{:}x]\,\bot \wedge$
$\bigwedge_{C\subset B} [B{:}leg{:}x]\,\bot \wedge [C{:}leg{:}x]\,\varphi)$
From def. $Leg_x\varphi$
2. $\bigvee_{C\in 2^{AGT\star}} (\neg\,[C{:}leg{:}x]\,\bot \wedge \bigwedge_{C\subset B} [B{:}leg{:}x]\,\bot \wedge$
$[C{:}leg{:}x]\,\varphi) \rightarrow \bigvee_{C\in 2^{AGT\star}} (\neg\,[C{:}leg{:}x]\,\bot \wedge$
$\bigwedge_{C\subset B} [B{:}leg{:}x]\,\bot \wedge [C{:}leg{:}x]\,\neg\,[C{:}leg{:}x]\,\bot \wedge$
$\bigwedge_{C\subset B} [C{:}leg{:}x]\,[B{:}leg{:}x]\,\bot \wedge [C{:}leg{:}x]\,[C{:}leg{:}x]\,\varphi)$
From Axiom $\mathbf{4_{Accept}}$ and Axiom $\mathbf{RolePlay}$
3. $(\bigwedge_{1\leq i\leq n} [C{:}leg{:}x]\,\varphi_i) \leftrightarrow ([C{:}leg{:}x] \bigwedge_{1\leq i\leq n} \varphi_i)$
Standard principle of the normal modal operator $[C{:}leg{:}x]$
4. $\bigvee_{C\in 2^{AGT\star}} (\neg\,[C{:}leg{:}x]\,\bot \wedge \bigwedge_{C\subset B} [B{:}leg{:}x]\,\bot \wedge$
$[C{:}leg{:}x]\,\neg\,[C{:}leg{:}x]\,\bot \wedge$
$\bigwedge_{C\subset B} [C{:}leg{:}x]\,[B{:}leg{:}x]\,\bot \wedge [C{:}leg{:}x]\,[C{:}leg{:}x]\,\varphi) \rightarrow$
$\bigvee_{C\in 2^{AGT\star}} (\neg\,[C{:}leg{:}x]\,\bot \wedge \bigwedge_{C\subset B} [B{:}leg{:}x]\,\bot \wedge$
$[C{:}leg{:}x]\,\neg\,[C{:}leg{:}x]\,\bot \wedge$
$[C{:}leg{:}x] \bigwedge_{C\subset B} [B{:}leg{:}x]\,\bot \wedge [C{:}leg{:}x]\,[C{:}leg{:}x]\,\varphi)$
From 3
5. $\bigvee_{C\in 2^{AGT\star}} (\neg\,[C{:}leg{:}x]\,\bot \wedge \bigwedge_{C\subset B} [B{:}leg{:}x]\,\bot \wedge$
$[C{:}leg{:}x]\,\neg\,[C{:}leg{:}x]\,\bot \wedge$
$[C{:}leg{:}x] \bigwedge_{C\subset B} [B{:}leg{:}x]\,\bot \wedge [C{:}leg{:}x]\,[C{:}leg{:}x]\,\varphi) \rightarrow$
$\bigvee_{C\in 2^{AGT\star}} (\neg\,[C{:}leg{:}x]\,\bot \wedge \bigwedge_{C\subset B} [B{:}leg{:}x]\,\bot \wedge$
$[C{:}leg{:}x]\,(\neg\,[C{:}leg{:}x]\,\bot \wedge \bigwedge_{C\subset B} [B{:}leg{:}x]\,\bot \wedge$
$[C{:}leg{:}x]\,\varphi))$
From 3
6. $\bigvee_{C\in 2^{AGT\star}} (\neg\,[C{:}leg{:}x]\,\bot \wedge \bigwedge_{C\subset B} [B{:}leg{:}x]\,\bot \wedge$
$[C{:}leg{:}x]\,(\neg\,[C{:}leg{:}x]\,\bot \wedge \bigwedge_{C\subset B} [B{:}leg{:}x]\,\bot \wedge$
$[C{:}leg{:}x]\,\varphi)) \rightarrow$
$\bigvee_{C\in 2^{AGT\star}} (\neg\,[C{:}leg{:}x]\,\bot \wedge \bigwedge_{C\subset B} [B{:}leg{:}x]\,\bot \wedge$
$[C{:}leg{:}x]\,Leg_x\varphi)$
From def. $Leg_x\varphi$
7. $\bigvee_{C\in 2^{AGT\star}} (\neg\,[C{:}leg{:}x]\,\bot \wedge \bigwedge_{C\subset B} [B{:}leg{:}x]\,\bot \wedge$
$[C{:}leg{:}x]\,Leg_x\varphi) \rightarrow Leg_xLeg_x\varphi$
From def. $Leg_xLeg_x\varphi$
8. $Leg_x\varphi \rightarrow Leg_xLeg_x\varphi$
From 1-2, 4-7

Proof of Theorem 3.

$$\vdash_{\mathcal{L}} Leg_x P_x\varphi \rightarrow P_x\varphi$$

1. $Leg_x P_x\varphi \rightarrow Leg_x \neg O_x \neg\varphi$
   From def. $P_x\varphi$
2. $\neg(Leg_x \neg O_x \neg\varphi \wedge Leg_x O_x \neg\varphi)$
   From Theorem 2e
3. $Leg_x \neg O_x \neg\varphi \rightarrow \neg Leg_x O_x \neg\varphi$
   From 2
4. $Leg_x P_x\varphi \rightarrow \neg Leg_x O_x \neg\varphi$
   From 1,3
5. $\neg Leg_x O_x \neg\varphi \leftrightarrow \neg O_x \neg\varphi$
   From Axiom **LegPower**
6. $Leg_x P_x\varphi \rightarrow P_x\varphi$
   From 4,5 and def. $P_x\varphi$

Proof of Theorem 5.

$$\vdash_{\mathcal{L}} (ExPower(i,a,\alpha,\varphi,x) \wedge NoChange(x, i{:}\alpha)) \rightarrow$$

$$After_{i:\alpha} Leg_x\varphi$$

1. $(ExPower(i,a,\alpha,\varphi,x) \wedge NoChange(x, i{:}\alpha)) \rightarrow$
   $(Power(a,\alpha,\varphi,x) \wedge Leg_x(\neg\,[i{:}a{:}x]\,\bot \wedge Happens_{i:\alpha}\top) \wedge$
   $NoChange(x, i{:}\alpha))$
   From def. $ExPower(i,a,\alpha,\varphi,x)$
2. $(Power(a,\alpha,\varphi,x) \wedge Leg_x(\neg\,[i{:}a{:}x]\,\bot \wedge Happens_{i:\alpha}\top) \wedge$
   $NoChange(x, i{:}\alpha)) \rightarrow$
   $((\neg\,[i{:}a{:}x]\,\bot \overset{x}{\rhd} After_{i:\alpha}\varphi) \wedge Leg_x(\neg\,[i{:}a{:}x]\,\bot \wedge$
   $Happens_{i:\alpha}\top) \wedge NoChange(x, i{:}\alpha))$
   From def. $Power(a,\alpha,\varphi,x)$
3. $((\neg\,[i{:}a{:}x]\,\bot \overset{x}{\rhd} After_{i:\alpha}\varphi) \wedge Leg_x(\neg\,[i{:}a{:}x]\,\bot \wedge$
   $Happens_{i:\alpha}\top) \wedge NoChange(x, i{:}\alpha)) \rightarrow$
   $(Leg_x(\neg\,[i{:}a{:}x]\,\bot \rightarrow After_{i:\alpha}\varphi) \wedge Leg_x(\neg\,[i{:}a{:}x]\,\bot \wedge$
   $Happens_{i:\alpha}\top) \wedge NoChange(x, i{:}\alpha))$
   From def. $\neg\,[i{:}a{:}x]\,\bot \overset{x}{\rhd} After_{i:\alpha}\varphi$
4. $(Leg_x(\neg\,[i{:}a{:}x]\,\bot \rightarrow After_{i:\alpha}\varphi) \wedge Leg_x(\neg\,[i{:}a{:}x]\,\bot \wedge$
   $Happens_{i:\alpha}\top) \wedge NoChange(x, i{:}\alpha)) \rightarrow$
   $(Leg_x(\neg\,[i{:}a{:}x]\,\bot \rightarrow After_{i:\alpha}\varphi) \wedge Leg_x \neg\,[i{:}a{:}x]\,\bot \wedge$
   $Leg_x Happens_{i:\alpha}\top \wedge NoChange(x, i{:}\alpha))$
   From standard principles of the normal modal operator Leg_x (*i.e.* $Leg_x(\varphi \wedge \psi) \leftrightarrow (Leg_x\varphi \wedge Leg_x\psi)$ is a theorem of $\mathcal{L}$)
5. $(Leg_x(\neg\,[i{:}a{:}x]\,\bot \rightarrow After_{i:\alpha}\varphi) \wedge Leg_x \neg\,[i{:}a{:}x]\,\bot \wedge$
   $Leg_x Happens_{i:\alpha}\top \wedge NoChange(x, i{:}\alpha)) \rightarrow$
   $(Leg_x After_{i:\alpha}\varphi \wedge Leg_x Happens_{i:\alpha}\top \wedge$
   $NoChange(x, i{:}\alpha))$
   From Theorem 2d
6. $(Leg_x After_{i:\alpha}\varphi \wedge Leg_x Happens_{i:\alpha}\top \wedge$
   $NoChange(x, i{:}\alpha)) \rightarrow$
   $(Leg_x(After_{i:\alpha}\varphi \wedge Happens_{i:\alpha}\top) \wedge NoChange(x, i{:}\alpha))$
   From standard principles of the normal modal operator Leg_x
7. $(Leg_x(After_{i:\alpha}\varphi \wedge Happens_{i:\alpha}\top) \wedge$
   $NoChange(x, i{:}\alpha)) \rightarrow$
   $(NoChange(x, i{:}\alpha) \wedge \bigvee_{C\in 2^{AGT\star}}(Leg(C,x) \wedge$
   $[C{:}leg{:}x]\,(After_{i:\alpha}\varphi \wedge Happens_{i:\alpha}\top)))$
   From def. $Leg_x(After_{i:\alpha}\varphi \wedge Happens_{i:\alpha}\top)$
8. $(NoChange(x, i{:}\alpha) \wedge \bigvee_{C\in 2^{AGT\star}}(Leg(C,x) \wedge$
   $[C{:}leg{:}x]\,(After_{i:\alpha}\varphi \wedge Happens_{i:\alpha}\top))) \rightarrow$
   $(NoChange(x, i{:}\alpha) \wedge \bigvee_{C\in 2^{AGT\star}}(Leg(C,x) \wedge$
   $[C{:}leg{:}x]\,After_{i:\alpha}\varphi \wedge [C{:}leg{:}x]\,Happens_{i:\alpha}\top))$
   From standard principles of the normal modal operator $[C{:}leg{:}x]$
9. $(NoChange(x, i{:}\alpha) \wedge \bigvee_{C\in 2^{AGT\star}}(Leg(C,x) \wedge$
   $[C{:}leg{:}x]\,After_{i:\alpha}\varphi \wedge [C{:}leg{:}x]\,Happens_{i:\alpha}\top)) \rightarrow$
   $(NoChange(x, i{:}\alpha) \wedge \bigvee_{C\in 2^{AGT\star}}(Leg(C,x) \wedge$
   $\neg\,[C{:}leg{:}x]\,\bot \wedge [C{:}leg{:}x]\,After_{i:\alpha}\varphi \wedge$
   $[C{:}leg{:}x]\,Happens_{i:\alpha}\top))$
   From def. $Leg(C,x)$
10. $(NoChange(x, i{:}\alpha) \wedge \bigvee_{C\in 2^{AGT\star}}(Leg(C,x) \wedge$
    $\neg\,[C{:}leg{:}x]\,\bot \wedge [C{:}leg{:}x]\,After_{i:\alpha}\varphi \wedge$
    $[C{:}leg{:}x]\,Happens_{i:\alpha}\top)) \rightarrow$
    $(NoChange(x, i{:}\alpha) \wedge \bigvee_{C\in 2^{AGT\star}}(Leg(C,x) \wedge$
    $\neg\,[C{:}leg{:}x]\,After_{i:\alpha}\bot \wedge [C{:}leg{:}x]\,After_{i:\alpha}\varphi \wedge))$
    From def. $Happens_{i:\alpha}\top$ and standard principles of the operator $[C{:}leg{:}x]$ (*i.e.* $\neg\,[C{:}leg{:}x]\,\bot \wedge [C{:}leg{:}x]\,Happens_{i:\alpha}\top$ implies $\neg\,[C{:}leg{:}x]\,After_{i:\alpha}\bot$)
11. $(NoChange(x, i{:}\alpha) \wedge \bigvee_{C\in 2^{AGT\star}}(Leg(C,x) \wedge$
    $\neg\,[C{:}leg{:}x]\,After_{i:\alpha}\bot \wedge [C{:}leg{:}x]\,After_{i:\alpha}\varphi)) \rightarrow$
    $(NoChange(x, i{:}\alpha) \wedge \bigvee_{C\in 2^{AGT\star}}(Leg(C,x) \wedge$
    $After_{i:\alpha}\,[C{:}leg{:}x]\,\varphi))$
    From Axiom **NF**
12. $(NoChange(x, i{:}\alpha) \wedge \bigvee_{C\in 2^{AGT\star}}(Leg(C,x) \wedge$
    $After_{i:\alpha}\,[C{:}leg{:}x]\,\varphi)) \rightarrow$
    $(\bigwedge_{C\in 2^{AGT\star}}(Leg(C,x) \rightarrow After_{i:\alpha}Leg(C,x)) \wedge$
    $\bigvee_{C\in 2^{AGT\star}}(Leg(C,x) \wedge After_{i:\alpha}\,[C{:}leg{:}x]\,\varphi))$
    From def. $NoChange(x, i{:}\alpha)$
13. $(\bigwedge_{C\in 2^{AGT\star}}(Leg(C,x) \rightarrow After_{i:\alpha}Leg(C,x)) \wedge$
    $\bigvee_{C\in 2^{AGT\star}}(Leg(C,x) \wedge After_{i:\alpha}\,[C{:}leg{:}x]\,\varphi)) \rightarrow$
    $\bigvee_{C\in 2^{AGT\star}}(After_{i:\alpha}Leg(C,x) \wedge After_{i:\alpha}\,[C{:}leg{:}x]\,\varphi)$
14. $\bigvee_{C\in 2^{AGT\star}}(After_{i:\alpha}Leg(C,x) \wedge$
    $After_{i:\alpha}\,[C{:}leg{:}x]\,\varphi) \rightarrow$
    $\bigvee_{C\in 2^{AGT\star}}(After_{i:\alpha}(Leg(C,x) \wedge [C{:}leg{:}x]\,\varphi))$
    From standard principles of the normal modal operator $After_{i:\alpha}$
15. $\bigvee_{C\in 2^{AGT\star}}(After_{i:\alpha}(Leg(C,x) \wedge [C{:}leg{:}x]\,\varphi)) \rightarrow$
    $After_{i:\alpha}(\bigvee_{C\in 2^{AGT\star}}(Leg(C,x) \wedge [C{:}leg{:}x]\,\varphi))$
    From standard principles of the normal modal operator $After_{i:\alpha}$
16. $After_{i:\alpha}(\bigvee_{C\in 2^{AGT\star}}(Leg(C,x) \wedge [C{:}leg{:}x]\,\varphi)) \rightarrow$
    $After_{i:\alpha}Leg_x\varphi$
    From def. $Leg_x\varphi$
17. $(ExPower(i,a,\alpha,\varphi,x) \wedge NoChange(x, i{:}\alpha)) \rightarrow$
    $After_{i:\alpha}Leg_x\varphi$
    From 1-16

References

Ågotnes, T.; van der Hoek, W.; Rodriguez-Aguilar, J.; Sierra, C.; and Wooldridge, M. 2007. On the logic of normative systems. In *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI'07)*, 1181–1186. AAAI Press.

Alchourrón, C., and Bulygin, E. 1971. *Normative systems*. New York: Springer.

Blackburn, P.; de Rijke, M.; and Venema, Y. 2001. *Modal Logic*. Cambridge: Cambridge University Press.

Boella, G., and van der Torre, L. 2003. Permissions and obligations in hierarchical normative systems. In *Proceedings of the Eighth International Conference on Artificial Intelligence and Law (ICAIL'03)*, 109–118. ACM Press.

Boella, G., and van der Torre, L. 2004. Regulative and constitutive norms in normative multiagent systems. In *Proceedings of the Ninth International Conference on Principles on Principles of Knowledge Representation and Reasoning(KR 2004)*, 255–266. AAAI Press.

Boella, G., and Van der Torre, L. 2007. Norm negotiation in multiagent systems. *International Journal of Cooperative Information Systems* 16(1):97–122.

Bratman, M. E. 1992. Practical reasoning and acceptance in context. *Mind* 101(401):1–15.

Bulygin, E. 1992. On norms of competence. *Law and Philosophy* 11(3):201–216.

Chellas, B. F. 1980. *Modal logic: an introduction*. Cambridge: Cambridge University Press.

Cohen, P. R., and Levesque, H. J. 1990. Intention is choice with commitment. *Artificial Intelligence* 42:213–261.

Cohen, L. J. 1992. *An essay on belief and acceptance*. New York, USA: Oxford University Press.

Conte, R., and Dignum, F. 2001. From social monitoring to normative influence. *The Journal of Artificial Societies and Social Simulation* 4(2).

Conte, R.; Castelfranchi, C.; and Dignum, F. 1998. Autonomous norm acceptance. In *Intelligent Agents V (ATAL'98)*, volume 1555 of *LNCS*, 99 – 112. Berlin: Springer Verlag.

Dignum, V., and Dignum, F. 2001. Modelling agent societies: Coordination frameworks and institutions. In Brazdil, P., and Jorge, A., eds., *Proceedings of the Tenth Portuguese Conference in Artificial Intelligence (EPIA'01)*, volume 2258 of *LNAI*. Berlin: Springer-Verlag.

Esteva, M.; Padget, J.; and Sierra, C. 2001. Formalizing a language for institutions and norms. In *Intelligent Agents VIII (ATAL'01)*, volume 2333 of *LNAI*, 348 – 366. Berlin: Springer Verlag.

Fagin, R.; Halpern, J.; Moses, Y.; and Vardi, M. 1995. *Reasoning about Knowledge*. Cambridge: MIT Press.

Gaudou, B.; Longin, D.; Lorini, E.; and Tummolini, L. 2008. Anchoring Institutions in Agents' Attitudes: Towards a Logical Framework for Autonomous MAS. In *Proceedings of the Seventh International Joint Conference on Autonomous Agents & Multi-Agent Systems (AAMAS'08)*, 728–735. ACM Press.

Gelati, J.; Rotolo, A.; Sartor, G.; and Governatori, G. 2004. Normative autonomy and normative co-ordination: Declarative power, representation, and mandate. *Artificial Intelligence and Law* 12(1-2):53–81.

Gerbrandy, J. 1999. *Bisimulations on Planet Kripke*. The Netherlands: PhD thesis, University of Amsterdam.

Gilbert, M. 1989. *On Social Facts*. London and New York: Routledge.

Grossi, D.; Meyer, J.-J. C.; and Dignum, F. 2006. Classificatory aspects of counts-as: An analysis in modal logic. *Journal of Logic and Computation* 16(5):613–643.

Hakli, P. 2006. Group beliefs and the distinction between belief and acceptance. *Cognitive Systems Research* 7:286–297.

Harel, D.; Kozen, D.; and Tiuryn, J. 2000. *Dynamic Logic*. Cambridge: MIT Press.

Hart, H. L. A. 1992. *The concept of law*. Oxford: Clarendon Press. new edition.

Herzig, A.; Lang, J.; and Polacsek, T. 2000. A modal logic for epistemic tests. In *Proceedings of the Fourteenth European Conference on Artificial Intelligence (ECAI 2000)*, 553–557. Berlin: IOS Press.

Jones, A., and Sergot, M. J. 1996. A formal characterization of institutionalised power. *Journal of the IGPL* 4:429–445.

Lopez y Lopez, F.; Luck, M.; and d'Inverno, M. 2004. Normative agent reasoning in dynamic societies. In *Proceedings of the Third International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS'04)*, 732 – 739. ACM Press.

Mantzavinos, C.; North, D.; and Shariq, S. 2004. Learning, institutions, and economic performance. *Perspectives on Politics* 2:75–84.

North, D. 1990. *Institutions, Institutional Change, and Economic Performance*. Cambridge: Cambridge University Press.

Rawls, J. 1955. Two concepts of rules. *The Philosophical Review* 64:3–32.

Raz, J. 1975. *Practical reason and norms*. London: Hutchinson.

Scherl, R. B., and Levesque, H. 2003. Knowledge, action, and the frame problem. *Artificial Intelligence* 144:1–39.

Searle, J. R. 1969. *Speech acts: An essay in the philosophy of language*. New York: Cambridge University Press.

Searle, J. R. 1995. *The Construction of Social Reality*. New York: The Free Press.

Stalnaker, R. 1984. *Inquiry*. Cambridge: MIT Press.

Tuomela, R. 1992. Group beliefs. *Synthese* 91:285–318.

Tuomela, R. 2007. *The Philosophy of Sociality*. Oxford: Oxford University Press.

Von Wright, G. H. 1963. *Norm and Action*. London: Routledge and Kegan.

4.3 Emotion and trust

A logical formalization of the OCC theory of emotions

C. Adam carole.adam.rmit@gmail.com
RMIT University, Melbourne, VIC, Australia

A. Herzig Andreas.Herzig@irit.fr
D. Longin Dominique.Longin@irit.fr
IRIT, Université de Toulouse, CNRS, Toulouse, France

December 16, 2021

Abstract

In this paper, we provide a logical formalization of the emotion triggering process and of its relationship with mental attitudes, as described in Ortony, Clore, and Collins's theory. We argue that modal logics are particularly adapted to represent agents' mental attitudes and to reason about them, and use a specific modal logic that we call Logic of Emotions in order to provide logical definitions of all but two of their 22 emotions. While these definitions may be subject to debate, we show that they allow to reason about emotions and to draw interesting conclusions from the theory.

Keywords: modal logics, BDI agents, emotions, OCC theory.

1 Introduction

There is a great amount of work concerning emotions in various disciplines such as philosophy [34, 79], economy [28, 52], neuroscience and psychology. In neuroscience, experiments have highlighted that individuals who do not feel emotions e.g. due to brain damage are unable to make rational decisions (see [20] for instance), refuting the commonsensical assumption that emotions prevent agents from being rational. Psychology provides elaborated theories of emotions ranging from their classification [25, 21] to their triggering conditions [48, 62] and their impact on various cognitive processes [31].

Computer scientists investigate the expression and recognition of emotion in order to design anthropomorphic systems that can interact with human users in a multi-modal way. Such systems are justified by the various forms of 'anthropomorphic behavior' that users ascribe to artifacts. This has lead to an increasing interest in Affective Computing, with particular focus on embodied agents [23], ambient intelligence [9], intelligent agents [81], *etc*. All these approaches generally aim at giving computers extended

capacities for enhanced functionality or more credibility. Intelligent embodied conversational agents (ECAs) use a model of emotions both to simulate the user's emotion and to show their affective state and personality. Bates has argued for the importance of emotions to make artificial agents more *believable*: "*It does not mean an honest or reliable character, but one that provides the illusion of life, and thus permits the audience's suspension of disbelief.*" [10, p. 122]. Indeed, there are many pieces of evidence suggesting that virtual agents and robots (interacting with humans) that are capable to display emotions, to recognize the human users' emotions, and to respond to their emotions in an appropriate way, allow to induce positive feelings in the humans during the interaction and to improve their performance. For instance it has been shown that emotions affect learning [12], so many computer scientists have added human-provided emotional scaffolding to their computer tutoring systems in order to increase both student persistence and commitment [5] and to improve learning [27]. In the same way, other researches show that machines which express emotions and provide emotional feedback to the user, allow to enhance the user's enjoyment [9, 66], her engagement [44] and performance in task achievement [63], her perception of the machine [14, 65] and can engage in more natural dialogs with her [11].

The great majority of these works are founded on psychological works about emotion. "What is the best theory of emotion today?" is a question where currently there is no consensus. A theory widely used by computer scientists is the one proposed by Ortony, Clore, and Collins (OCC henceforth). A reason is that this theory is relatively understandable by computer scientists, because it is founded on a combinatory approach of a finite set of criteria allowing to characterize emotions. (We are going to present this theory in more detail in Section 2.2, and are going to present more arguments in Section 2.4.)

OCC theory provides what may be called a semi-formal description language of emotion types. It neither accounts for relationships between the different components of emotions nor relationships between agents' emotions and their actions. The aim of this paper is to fill this gap by formalizing OCC theory with the help of a language describing agents' mental attitudes such as beliefs, goals or desires. In this way we stay as close as possible to the original psychological theory. More precisely, we aim at modelling the triggering process of emotions in intelligent agents endowed with mental states (*viz*. a set of mental attitudes about some contents). What we do is to describe how a given mental state contributes to the triggering of a given emotion. This problem has to be solved before formalizing the subsequent influence of emotions on any mental process and in particular on planning. In this paper we therefore focus on the influence of mental states on emotions, and do not address the influence of emotions on mental states.

Our aim is to model emotion in a logic of mental attitudes. Formal logic provides a universal vocabulary with a clear semantics and it allows reasoning, planning and explanation of an agent's behavior. A given formal definition of emotions may be criticized, but it still has the advantage to be unambiguous and to allow analysis and verification. In particular, all logical consequences of formal principles must remain intuitive: a logical formalization may reveal consequences (and even inconsistencies) that were 'hidden' in the theory and did not appear before. Formal definitions clearly articulate assumptions and allow to formally derive consequences of certain assump-

tions: they allow to clearly and concisely articulate the assumptions of a theory and to readily uncover the consequences. All in all, logical formalization is a well-defined scientific program to move forward and develop more widely accepted and clearly defined models.

The logic used here is a particular modal logic that grounds on the philosophy of language, of mind, and of action [13, 75, 76], and proposes to model agents *via* some key concepts such as *mental attitudes* (belief, preference, desirability), action and time. This framework is very close to those commonly used in the agent community and offers well-known interesting features: great explanatory power, formal verifiability, and a rigorous and well-established theoretical frame (from the point of view of both philosophy and formal logic). Note that we are not concerned at this stage with optimizations of our logical theory in view of particular applications; for the time being we leave this to agent designers who might use our model as a basis for their work.

Our aim is also to model emotion in a way that is as faithful as possible to psychology. Thus we believe that our logical theory is built on solid grounds given that OCC theory is a well-established psychological theory. The properties of our logic may be evaluated with respect to the following criteria : 1) the number and types of the emotions that are covered; 2) the examples given by psychologists (is our formalism able to account for these examples?); 3) the theorems following from our model (are these theorems intuitive and relevant? are they in accordance with the formalized psychological theory or do they run counter to it? *etc*.).

We also believe that the other way round our logic, thanks to its faithfulness to the OCC theory, may contribute to the assessment of this theory. For example the consistency of our logic demonstrates that OCC theory is free of contradictions.

In the rest of the paper, we expose the OCC theory underlying our work (Section 2). In Section 3 we introduce our logical framework. In Section 4 and Section 5 we detail the event-based and agent-based branches of the OCC theory and their formalization. In Section 6 we expose some theorems concerning emotions, particularly relating to causal and temporal links between them. In order not to overload the paper, the proofs of these theorems are gathered in the appendix. In Section 7 we discuss some existing logical models of emotions.

2 Emotion theories

To ensure the accuracy of a computational model of emotions, it is important to start from acknowledged psychological theories. There exist several kinds of psychological models of emotions: *evolutionist models* (*e.g*. [21]) that are mainly descriptive, giving taxonomies of basic emotions; *dimensional models* (*e.g*. [71]) that assume that all emotions are similar phenomena, only varying on the values of some dimensions like valence or arousal; these models were sometimes used to describe the dynamics of the expression of emotions (*e.g*. [11]); *cognitive appraisal theories* (*e.g*. [62]) that focus on the cognitive determination of emotions and on their adaptive function.

The concept of appraisal was first introduced by [8] to describe the triggering of emotions, together with the concept of action tendencies describing their effects. These two concepts were then studied in many approaches; we present some of the most im-

portant ones here (Section 2.3), in particular that of OCC theory (Section 2.2). Before that, let us first shortly speak about relationships between emotion and cognition through the concept of Intentionality (Section 2.1).

2.1 Emotion, cognition, Intentionality, and logic

The use of logic for emotion formalization may appear surprising at first glance, and one might consider that they cannot be married. Nevertheless, today the great majority of psychologists work with approaches where emotion and cognition are strongly connected (see [47] for instance: for him, emotion is a part of cognition). Logic can deal with cognition through the well-known BDI logics (for belief, desires, intentions). Cognition refers, among other things, to mental states and reasoning about them. Thus, to say that emotion is in cognition means that emotion is concerned with mental states.

In our view, emotions are always about a state of affairs of the world. In other words, emotion is an *Intentional* components of our mind in the sense of [76, p. 1]: "*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*". (Note that *intention* is just a particular form of Intentionality. To avoid confusions and following [76], we write "Intentionality" with an upper case letter.) For instance, belief and preference are among the Intentional mental states. Note that only some, but not all, mental states have Intentionality: for instance, forms of nervousness, elation, and undirected anxiety, that are diffuse without any clear link with an object of the world, are not Intentional states. Searle ([76, pp. 29–36]) has described some emotions as complex mental attitudes that can be expressed as a combination of beliefs and desires. We want to generalize this approach by applying it to OCC theory. In this perspective, the description of an emotion as a combination of beliefs and desires presupposes that the emotion under concern is an Intentional mental attitude. We therefore do not deal in this paper with other emotional states that are closer to *mood* or that are not *Intentional emotions*, in the sense that they are not concerned with or based on Intentional mental attitudes.

Following the great majority of psychologists, another difference between emotion and mood is that emotion has a very short duration in time. (See for instance [62, 48].) Thus, we can expect that an affective state having a long duration is not so much emotion as mood.

Finally, Intentional mental attitudes can be either conscious or unconscious, and this property is not related to Intentionality. (Searle shows that one can be conscious of non Intentional mental states, and conversely one can be unconscious of an Intentional state, see [76, Chap. 1].) Figure 1 pictures the situation.

Ortony and colleagues agree that "individual emotions can be specified in terms of personal or interpersonal situation descriptions that are sufficient to produce them" [62, p. 3]. More precisely they assume that "if the described situation contains the eliciting condition for a particular emotion, the experience of that emotion can be inferred" [62, p. 3]. This is clearly a cognitive approach where emotions are Intentional concepts. This supports our fundamental design choice: emotions can only occur in particular mental states formalized through the logical definition of the mental attitudes constituting their elicitation conditions. Our choice of a logic of mental attitudes is therefore justified both by the fact that it is an appropriate formalization of mental states (see

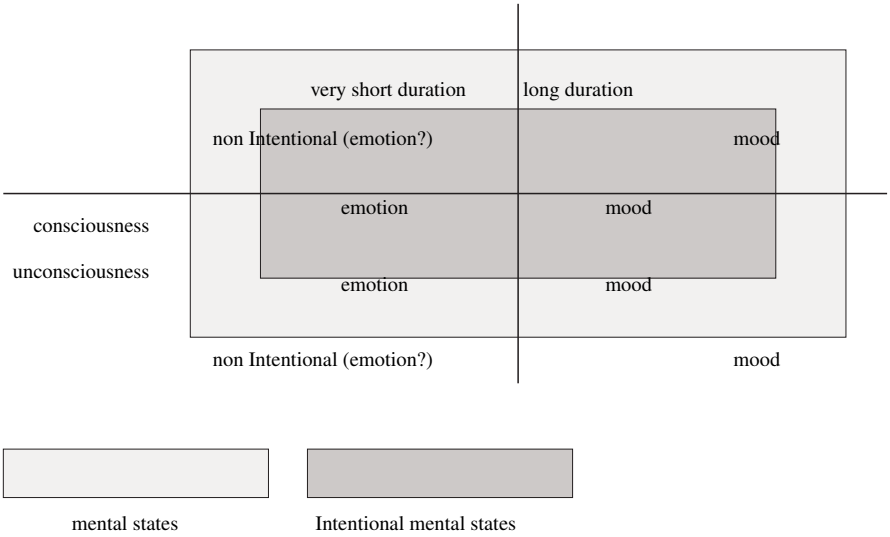

Figure 1: Emotions, mood, Intentional states, and duration

[19, 67, 68]), [72], [41] for instance) and by the fact that mental attitudes allow to express emotions.

2.2 Ortony, Clore and Collins's theory of emotion

[62] propose a cognitive appraisal theory that is structured as a three-branch typology, corresponding to three kinds of stimuli: consequences of events, actions of agents, and aspects of objects. Each kind of stimulus is appraised w.r.t. one central criterion, called *central appraisal variable*. An individual judges the following:

- the desirability of an event, *viz*. the congruence of its consequences with the individual's goals (an event is pleasant if it helps the individual to reach his goal, and unpleasant if it prevents him from reaching his goal);
- the approbation of an action, *viz*. its conformity to norms and standards;
- the attraction of an object, *viz*. the correspondence of its aspects with the individual's likings.

There are some secondary appraisal variables influencing the intensity of the generated emotion, such as the probability of an event, the degree of responsibility of the author of an action and the amount of effort that was provided.

The OCC typology contains twenty-two emotions types[1] that are grouped in six classes. The first branch contains three classes of emotions triggered by the appraisal

[1] According to the authors, an emotion type is "a distinct kind of emotion that can be realized in a variety of recognizably related forms" [62, p. 15], for example various intensities or various emphasis. In the sequel of this paper, to simplify the vocabulary we generally use the term "emotion" instead of "emotion type".

of the consequences of an event as to its desirability. *Well-being emotions* (joy, distress) arise when an individual appraises an event that has just occurred while only focusing on the desirability of its consequences for himself. *Fortunes-of-others emotions* (happy for, sorry for, resentment, gloating) arise when an individual appraises an event while focusing on its desirability for another individual. *Prospect-based emotions* such as hope or fear arise when an individual appraises the consequences of a prospected event (namely an event that has not occurred yet but is expected to do so) while focusing on the desirability of its consequences for himself. Other prospect-based emotions such as disappointment, relief, fears-confirmed, and satisfaction arise when an individual appraises an event that has just occurred and that was expected, while focusing on its desirability for himself.

The second branch contains only one class of emotion types (*Attribution emotions*) triggered by the appraisal of an action as to its approval, *viz*. its conformity to norms and standards. Thus, pride and shame arise when an individual appraises one of his own actions while focusing only on its approval ('does this action conform to the standards?') and not on its consequences. Admiration and reproach arise when an individual appraises an action of another individual while focusing only on its approval.

An other class, common both to Well-being emotions (first branch of the typology) and Attribution emotions (second branch of the typology) is *Compounds emotions (attribution-wellbeing)* (remorse, gratification, gratitude, anger) that arise when an individual appraises an action while focusing both on its approval an on the desirability of its consequences.

Here is a complex example where several of the above emotion types are involved. Suppose you and a friend of yours apply for the same position. You believe your CV is better, but then you learn that your friend got the position because he cheated a bit on his CV (say he over-emphasized his participation in some project and gave to some of his papers a "to appear" status although they are just submitted). According to OCC theory you might then feel (1) disappointed (confirmation-based), (2) happy for your friend (fortune of other), and (3) reproach (attribution emotion). The relative importance of these three emotions depends on the secondary appraisal variables, which is something we do not account for in our framework. What we deal with here is whether such emotions can indeed be triggered simultaneously by the same event, *viz*. whether the conjunction of these three emotions is consistent in our logic.

Finally, the third branch contains one class of emotions: *attraction emotions* (love, hate), triggered by the appraisal of the aspects of objects w.r.t. the individual's likings.

It is important to notice that the authors of the OCC theory intended it to be used in Artificial Intelligence:

> "(...) we would like to lay the foundation for a computationally tractable model of emotion. In other words, we would like an account of emotion that could in principle be used in an Artificial Intelligence (AI) system that would, for example, be able to reason about emotion."
>
> [62, p. 2]

This aim was pretty much reached since OCC theory is the most popular psychological model of emotions in computer science nowadays, and emotional agents widely employ it (*e.g*. [26], [69], [23], [43], [61]). However, it is not the only one, and we can

quote emotional agents based on Frijda's theory (*e.g.* [80]) as well as agents based on Lazarus's theory (*e.g.* [35]).

2.3 Other theories

[32] focuses on the action tendencies induced by emotions. A stimulus first passes through various steps of evaluation determining its characteristics: causes and consequences, relevance and congruence with interests, coping possibilities, urgency. Depending on the result, a control signal is generated to postpone or interrupt the current action. An action preparation is then created (action plan, action tendency, activation mode) that induces physiological changes, and finally an action is selected and executed. Frijda believes that it is the associated action tendency that differentiates basic emotions from each other. [22] build on this notion of action tendency to define the effect of four emotions on a rational agent's plans.

[48] presents a relational, motivational, cognitive theory of emotion. According to him, emotions result from the cognitive evaluation (or appraisal) of the interaction between an individual and its environment, w.r.t. his motivations and goals. Lazarus distinguishes between the primary appraisal, assessing the relevance and congruence of the stimulus w.r.t. the individual's well-being (that is, does the stimulus help or threaten one of the individual's goals?), and the secondary appraisal, evaluating the available resources to cope with the stimulus (can the individual do something to remove the threatening stimulus?). These two kinds of appraisal are not sequential: they can be executed in any order. Like Arnold, Lazarus considers that emotions induce action tendencies, that cause physiological modifications in order to help the individual adapting to his/her environment. Lazarus' theory is used in the EMA agent (*cf.* [35]) whose acronym is an homage to his book.

[74] considers emotions as a multicomponent process, with one cognitive component. He introduces an appraisal process consisting in a sequence of stimulus processing steps, called the *Stimulus Evaluation Checks*. This process sequentially evaluates the novelty and unexpectedness of the stimulus, its intrinsic agreeability, its congruence with the individual's goals, the coping possibilities, and its compatibility with norms. Contrarily to [48], these evaluations are ordered. Later, [73] associates to each emotion bodily responses, and in particular facial expressions in terms of *Action Units*. The latter are elements defined by [24] to represent the moves of the facial muscles. This theory is thus well-adapted to represent the dynamics of facial expressions of an animated agent (*e.g.* [36]).

2.4 Which theory to choose?

Appraisal theories importantly differ one from another on the appraisal criteria that are used, their order of application, and the precise definitions of emotions based on these criteria. We have chosen OCC theory because the careful study of this theory in comparison with others like Lazarus's indicated that it is better adapted to describe the emotions of a virtual agent for several reasons.

First, OCC theory is widely used in the design of emotional agents because its simplicity and implementability matches computer scientists' expectations and needs:

it seems that the combination of OCC's finite set of appraisal variables suffices for current applications.

Second, we completely agree that according to OCC, any emotion must be valenced and this valence must always be the same [62, pp. 29–32]. This excludes *de facto* states like surprise (that can be either good or bad, or even neither good nor bad) or "feeling abandoned" (in this state one can be sad, but can also not let this get one down and get one's hope up) from being emotions. (In some works, surprise is considered to be an emotion, see [78], [60], [53] or [54] for examples in recent works, or Ekman's works in 70th and [21] for older works.) Besides, the necessity for an emotion to be valenced has also the advantage to provide a clear test to differentiate emotions from close notions that are not valenced. Moreover, valence is something naturally captured by logic, making the OCC theory particularly well adapted for a logical formalization.

Third, OCC theory has a simple and elegant tree structure, and uses concepts that have been well-studied in logic such as beliefs, desires and standards. This makes the formalization task easier. In Searle's view [76, Section 1.5], every Intentional state is not reducible to belief and desire, but every Intentional state contains a belief, a desire, or both a belief and a desire. So-called BDI logics developed in the field of Artificial Intelligence in the last fifteen years (see [19, 67] for instance) offer expressive frameworks to represent agents' mental attitudes such as beliefs and desires (see [56, 41] and [55] for instance) and to reconstruct on their basis the cognitive layer of emotions (see [2, 3] and also [81] for instance).

Finally, OCC theory is quite exhaustive, which is important to design robust and versatile agents, *viz.* agents that can emotionally react to a great variety of situations. On the contrary Lazarus' theory is more precise but seems to be less exhaustive (see [2, Chapter 4] for a more detailed comparison). We believe that the logical formalization of both theories will allow to compare them in depth in a close future.

The next section presents the logical framework.

3 Logical framework: the EL logic

The formal framework of this article is based on our previous BDI framework [41]) that in turn is based on BDI logics (see for instance [19, 67, 72]). We minimally extend this standard framework by integrating OCC's appraisal variables. As we have said, we restrict our attention to emotion triggering conditions, disregarding the influence of emotions on beliefs, desires and intentions. OCC's emotion triggering conditions do not refer to the mental attitude of intention, that is therefore not required here. As we are only concerned with event-based emotions and agent-based emotions, we here only need to model the *desirability* and *praiseworthiness* variables of OCC theory. But let us first explain these variables and the choices we made in order to model them.

In OCC theory, desirability is about events and is close to the notion of utility. When an event occurs it can satisfy or interfere with agent's goals, and the desirability *variable* has therefore two *aspects*:

> "one corresponding only to the degree to which the event in question appears to have beneficial (*viz.* positively desirable) consequences, and the

other corresponding to the degree to which it is perceived as having harmful (*viz.* negatively desirable, or undesirable) consequences."

[62, p. 49]

It is thus a valenced variable, and an event can be at the same time desirable and undesirable (with respect to the agent's current goals). Desirability (*viz.* the positive *aspect* of the desirability *variable*) only influences positive emotions, whereas undesirability (*viz.* the negative *aspect* of the desirability *variable*) only influences negative emotions. It follows that the same event can trigger both positive and negative emotions.

While we agree with OCC theory that the primary event may be both desirable and undesirable[2], such a feature makes a logical formalization difficult because it requires either a paraconsistent notion of desirability such that "φ is desirable" and "$\neg\varphi$ is desirable" are consistent, or a binary notion of desirability that is relativized to goals. Both options would induce several difficulties, that would distract us from our aims; in particular there is no available logic of the latter binary desirability in the literature. A way out is to shift the focus from desirability of events to desirability of *consequences of events*: when someone says that an event is both desirable and undesirable, we are entitled to ask which aspect of this event is desirable and which is undesirable. When considering consequences of events rather than events themselves we may safely suppose that these consequences are either desirable or undesirable with respect to current goals, but not both. For instance, someone's death can entail both an affective loss (undesirable consequence) and the inheritance of a big amount of money (desirable consequence) [62]. In this example, clearly, the goal concerned with the desirability of the event (that is, to get a big amount of money, or to be rich) is different from the goal concerned with the undesirability of this event (that is, to keep a loved person alive). Correspondingly, our desirability operators have the following properties:

- desirability and undesirability are about consequences of events (and not about the event itself);
- an event can have several consequences;
- each of these consequences cannot be both desirable and undesirable;
- each of these consequences can be evaluated with respect to goals (that may be either achievement goals or maintenance goals).

We formalize the above example by saying that an emotional loss is undesirable and a big amount of money is desirable, while the friend's death is neither desirable nor undesirable.

Things are similar concerning OCC's *praiseworthiness* variable, which concerns the evaluation of actions performed by agents; this evaluation is with respect to standards and has two aspects: actions can be praiseworthy (when they conform to standards) or blameworthy (when they violate standards). Though, we want to avoid to analyze standards in more depth and do not describe how to construct the two aspects

[2]For instance the authors give the example of the death of one's friend suffering from a painful disease; on the one hand the loss of one's friend is undesirable, but on the other hand the end of his suffering is desirable.

of the praiseworthiness variable. We simply define two types of modal operators: one characterizing the praiseworthiness of consequences of actions, and one characterizing the blameworthiness of consequences of actions. Just as for desirability and undesirability, we will consider that such a consequence cannot be both praiseworthy and blameworthy at the same time.

3.1 Syntax

The syntactic primitives of our logic of emotions EL are as follows: a nonempty finite set of agents $AGT = \{i_1, i_2, \ldots i_n\}$, a nonempty finite set of atomic events $EVT = \{e_1, e_2, \ldots, e_p\}$, and a nonempty set of atomic propositions $ATM = \{p_1, p_2 \ldots\}$. The variables i, j, k... denote agents; α, β, γ... denote events; and p, q... denote propositional letters (atomic propositions). The expression $i{:}e$ represents an event e intentionally caused by agent i. We say that $i{:}e$ is an action that is performed by i.

The language $\mathcal{L}_{EL}$ of the EL logic is defined by the following BNF (Backus Naur Form):

$$\varphi ::= \bot \mid p \mid \neg\varphi \mid \varphi \vee \varphi \mid Bel_i\,\varphi \mid Prob_i\,\varphi \mid Des_i\,\varphi$$
$$\mid Idl\ \varphi \mid After_{i:\alpha}\,\varphi \mid Before_{i:\alpha}\,\varphi \mid G\varphi \mid H\varphi$$

where p rangers over ATM, i ranges over AGT and $i{:}\alpha$ ranges over $AGT \times EVT$. The classical boolean connectives $\wedge$ (conjunction), $\rightarrow$ (material implication), $\leftrightarrow$ (material equivalence) and $\top$ (tautology) are defined from $\neg$ (negation), $\vee$ (disjunction) and $\bot$ (contradiction) in the usual manner.

$Bel_i\,\varphi$ reads "agent i believes that φ is true". Belief is understood as subjective knowledge, alias truth in all worlds that are possible for the agent: i does not doubt. For instance, $Bel_{i_1}\ weatherNice$ represents the fact that, from i_1's point of view the weather is nice: i_1 has no doubt about the truth of this fact (but may be wrong).

$Prob_i\,\varphi$ reads "agent i believes that φ is more probable than $\neg\varphi$", or "i believes that φ probable" for short. This is a weaker form of belief than $Bel_i\,\varphi$. For example, if agent i_1 is still in bed, $Prob_{i_1}\ weatherNice$ means that i_1 believes that the weather is probably nice (but i_1 may not be sure about this). What an agent believes is necessarily probable for him, but not the other way round: when i_1 believes that p then p is probable for i_1. (We give more details in the sequel.) Several researchers have investigated logics of probability, mainly in a quantitative [29] or comparative way [77]. A few researchers studied a more qualitative notion of probability [15, 39], weak belief (Lenzen [50, 51]) or likelihood [38, 37]. All these are based on subjective probability measures. We adopt Burgess's logic, basically because we do not need numbers for our purposes (but they might be added later when investigating particular applications), and because it integrates smoothly with Hintikka's logic of belief that we are going to use.

$Des_i\,\varphi$ reads "φ is desirable for i". As we have motivated above, instead of desirability of events we here rather deal with desirability of consequences of events. These consequences are evaluated with respect to goals. According to the OCC theory, goals can be either achievement goals (the agent wants to achieve something that is not currently true) or maintenance goals (the agent wants to maintain something that is

already true). Moreover, we do not explain the relationship between goals and desirability because goals do not play an explicit role in our definition of emotions. However, goals can be constructed from what is desirable, and intentions can be constructed from goals. (See [41] and [17] for more details about such constructions.) In our view, every (achievement or maintenance) goal is about something that is desirable. Thus, if a consequence of an event is (a part of) a goal, then this consequence is desirable. Here, instead of an *occurrent mental attitude* we rather use the notion of *dispositional attitude*, that corresponds with Bratman's notion of desire and with Cohen&Levesque's notion of goal.[3]

$Idl\ \varphi$ reads "ideally, φ is true". The notion of ideality considered here is taken in a large sense: it embraces all the rules more or less strongly imposed by some authority. They can be strongly explicit (like laws) or more or less implicit (like social or moral rules). When $Idl\ \varphi$ is true then φ is a kind of social preference that is attached to the groups to which the agent belongs. They may therefore differ from the agent's personal preferences. $Idl\ driveRight$, for instance, means that ideally, one drives on the right side of the road, and $Idl\ helpSbInDistress$ means that ideally, one helps somebody in distress.

$After_{i:\alpha}\,\varphi$ reads "φ will be true after performance of action α by i". This operator allows to describe what is true after the execution of an action, in particular the effects of this action. For instance, $After_{i_1:raiseHand}\ rightToSpeak_{i_1}$ means that after agent i_1 has raised its hand (say in the classroom) it will have the right to speak. The fact that φ will be true after the performance of action α is conditional on the performance of α: it does not entail that α is currently executed, nor that i intends to execute it. $After_{i:\alpha}\,\bot$ reads "action α is not executed by agent i". For instance, $After_{i_2:drive}\,\bot$ means that agent i_2 is not going to drive in the current situation (for instance because i_2 does not have a car).

$Before_{i:\alpha}\,\varphi$ reads "φ was true before performance of action α by i". It is symmetric to $After_{i:\alpha}$ for the past. $Before_{i:\alpha}\,\bot$ means "i has not just executed action α". $Before_{i_2:holdsNut}\,holdsNut_{i_2}$, for instance, means that before crunching a nut, agent i_2 must hold a nut, and $Before_{i_1:drink}\,\bot$ means that drinking was not i_1's last action.

$G\varphi$ reads "henceforth φ is going to be true". The notion of time that we use here is linear time. It means that states of world are organized in a linear manner, in what is called "histories" in the litterature. Thus, $G\varphi$ means that φ is true on the current history from now and everywhere in the future. For instance, $GglassIsBroken$ means that the glass is henceforth broken.

$H\varphi$ reads "φ has always been true in the past". Thus, it means that φ is true on the current history everywhere in the past including now. For instance, $H\neg JohnIsDead$ means that until and including now, John is not dead.

For convenience, we also define the following abbreviations:

$$Happens_{i:\alpha}\,\varphi \stackrel{def}{=} \neg After_{i:\alpha}\,\neg\varphi \qquad (\text{Def}_{Happens_{i:\alpha}})$$

[3] Several concepts of desire exist in the literature. Desire is often viewed as an *occurrent mental attitude*: an attitude that holds here and now, and that is abandoned as soon as it is satisfied, such as an agent's desire on a rainy day that the sun shines, which is dropped when finally the sun comes out. This is similar to Bratman's concept of intention [13] and to Cohen&Levesque's concept of achievement goal [19].

$$Done_{i:\alpha}\,\varphi \stackrel{def}{=} \neg Before_{i:\alpha}\,\neg\varphi \qquad (\text{Def}_{Done_{i:\alpha}})$$

$$F\varphi \stackrel{def}{=} \neg G\neg\varphi \qquad (\text{Def}_{F})$$

$$P\varphi \stackrel{def}{=} \neg H\neg\varphi \qquad (\text{Def}_{P})$$

$$Idl_i\,\varphi \stackrel{def}{=} Bel_i\,Idl\,\varphi \qquad (\text{Def}_{Idl_i})$$

$Happens_{i:\alpha}\,\varphi$ reads "α is about to be performed by agent i, after which φ will be true"[4]. In particular, $Happens_{i:\alpha}\,\top$ reads "action α is about to be performed by agent i". For instance, $Happens_{i_1:tossCoin}\,(heads \vee tails)$ means i_1 is about to toss a coin, after which the coin will be either heads or tails.

$Done_{i:\alpha}\,\varphi$ reads "α has just been performed by agent i, and φ was true before" and $Done_{i:\alpha}\,\top$ reads that agent i has just performed action α. For instance, $Done_{i_2:toDrinkBeer}\,Done_{i_1:toDrinkCoke}\,\top$ means that agent i_2 has just drunk a beer and just before that, agent i_1 had drunk a coke.

$F\varphi$ reads "φ is true or will be true at some future instant", and $P\varphi$ reads "φ is or was true". For example, $PsunIsShining \wedge FsunIsShining$ means that there is a past instant when the sun was shining and there is a future instant when the sun will be shining.

Finally, $Idl_i\,\varphi$ reads "from the point of view of the agent i, it is ideal that φ be true". It will be convenient to suppose that it represents an agent's moral norms, that is, the norms that the agent has internalized as true. For instance, $Idl_{i_1}\,beVegetarian$ means that for agent i_1 one should be vegetarian, and $Idl_{i_2}\,\neg(Drunk \wedge Driving)$ means that for agent i_2 it is unideal to drive drunk. Note that in principle not every known ideal (*viz.* $Bel_i\,Idl\,\varphi$) becomes an internalized ideal ($Idl_i\,\varphi$), i.e., the left to right implication $Bel_i\,Idl\,\varphi \rightarrow Idl_i\,\varphi$ of the Definition Def_{Idl_i} is not generally valid. The difference is subtle (see [1] for more details) and here we adopt this simplification because it allows us to avoid an investigation of the relation between internalized and non-internalized standards.

3.2 Semantics

We use a standard Kripke semantics in terms of possible worlds and accessibility relations. The less standard feature is a neighborhood function for the modal operator of probability.

3.2.1 *EL* frames

At the base of Kripke semantics there is a set of possible worlds W together with accessibility relations for every modal operator. While in most presentations an accessibility relation is a subset of the cartesian product $W \times W$, we here use an equivalent presentation in terms of mappings from W to 2^W.

An *EL* frame is a 7-tuple $\mathcal{F} = \langle W, \mathscr{B}, \mathscr{P}, \mathscr{D}, \mathscr{I}, \mathscr{A}, \mathscr{G}\rangle$ where:

- W is a set of possible worlds;

[4]Note that the operators *Happens* can be read in this way (which is not their standard dynamic logic reading) because we have supposed determinism: time is linear, entailing that if an action is feasible, then it will happen.

- $\mathscr{B} : AGT \longrightarrow (W \longrightarrow 2^W)$ is the accessibility relation that associates each agent $i \in AGT$ and possible world $w \in W$, with the set $\mathscr{B}_i(w)$ of possible worlds compatible with the beliefs of agent i in w;
- $\mathscr{P} : AGT \longrightarrow (W \longrightarrow 2^{2^W})$ is the function that associates each agent $i \in AGT$ and possible world $w \in W$ with a set of sets of possible worlds $\mathscr{P}_i(w)$ (the *neighborhoods* of w);
- $\mathscr{D} : AGT \longrightarrow (W \longrightarrow 2^W)$ associates each agent $i \in AGT$ and possible world $w \in W$ with the set $\mathscr{D}_i(w)$ of worlds compatible with what is desirable for the agent i in the world w;
- $\mathscr{I} : W \longrightarrow 2^W$ associates each possible world $w \in W$ with the set $\mathscr{I}(w)$ of ideal worlds;
- $\mathscr{A} : AGT \times ACT \longrightarrow (W \longrightarrow 2^W)$ associates each action $i{:}\alpha \in AGT \times ACT$ and possible world $w \in W$ with the set $\mathscr{A}_{i:\alpha}(w)$ of possible worlds resulting from the performance of α by agent i in w;
- $\mathscr{G} : W \longrightarrow 2^W$ associates each possible world $w \in W$ with the set $\mathscr{G}(w)$ of possible worlds in the future of w.

The set $\mathscr{B}_i(w)$ is called a belief state.

3.2.2 Semantical constraints

We impose to our frames the following semantical constraints.

All the accessibility relations $\mathscr{B}_i$ are serial, transitive and euclidian. **($\mathbf{SC_1}$)**

Thus, belief states are equivalence classes: an agent views several alternative worlds to the real world but cannot distinguish between each of these alternatives. Note that contrarily to knowledge the real world is not necessarily contained in an agent's belief state. Seriality ensures that beliefs are rational: an agent cannot simultaneously believe that p is true and that its negation $\neg p$ is true. Due to the transitivity and euclidianity of the $\mathscr{B}_i$ relations, agents are aware of their beliefs: if $w \in \mathscr{B}_i(w')$ then $\mathscr{B}_i(w) = \mathscr{B}_i(w')$.

If φ is probable for i (*viz.* φ is true in all the worlds of some neighborhood, see [18, chap. 7] for more details), then $\neg\varphi$ is not (since each other neighborhood contains at least one world where φ is true). This corresponds to the following constraint:

For every $w \in W$, if $U_1, U_2 \in \mathscr{P}_i(w)$ then $U_1 \cap U_2 \neq \emptyset$. **($\mathbf{SC_2}$)**

Moreover, in order to ensure that at least tautologies are probable, we impose that:

$\mathscr{P}_i(w) \neq \emptyset$ for every $w \in W$. **($\mathbf{SC_3}$)**

Finally, we impose that the neighborhoods in $\mathscr{P}_i(w)$ are subsets of the belief state:

$\forall U \in \mathscr{P}_i(w), U \neq \emptyset$ **($\mathbf{SC_4}$)**

which entails that belief implies probability.[5]

As explained in the previous section, when desirability is about propositions rather than actions, it is convenient to postulate to consider that desirability is rational: if a proposition is desirable then its converse is not desirable. This is imposed by the following semantical constraint:

$$\text{All the accessibility relations } \mathscr{D}_i \text{ are serial.} \qquad (\mathbf{SC_5})$$

The situation is similar for ideality: intuitively, the logic of ideality operators is the same as Standard Deontic Logic. (See [7] for more details about Deontic Logic.) Here, the rationality of ideals is justified by the fact that law, moral, habits, standards, *etc*. are in principle coherent. Thus, if something is ideally true, then its converse must not be true.

$$\text{All the accessibility relations } \mathscr{I} \text{ are serial.} \qquad (\mathbf{SC_6})$$

Concerning action, we impose that for every $w \in W$:

$$\text{If } w' \in \mathscr{A}_\alpha(w) \text{ and } w'' \in \mathscr{A}_\beta(w) \text{ then } w' = w''. \qquad (\mathbf{SC_7})$$
$$\text{If } w \in \mathscr{A}_\alpha(w') \text{ and } w \in \mathscr{A}_\beta(w'') \text{ then } w' = w''. \qquad (\mathbf{SC_8})$$

First, this imposes that actions are organized into histories. It does not impede the parallel execution of several actions, but it guarantees that all these parallel actions lead to the same world, i.e., the same time point in the same history. It imposes that all the actions take place in the same history, where the outcome world is the same for all actions performed by all agents.[6] Second, these constraints impose that actions take one time step. Suppose that α and β are performed during the performance of γ. That is: $w' \in \mathscr{A}_\alpha(w)$, $w'' \in \mathscr{A}_\beta(w')$ and $w'' \in \mathscr{A}_\gamma(w)$ hold. Thus, (1) imposes that $w' = w''$, and (2) imposes that $w = w'$, which entails that $w = w' = w''$. Thus, in this case, all actions are reduced to the 'skip' action ('do nothing') and the world remains unchanged. Therefore, actions are deterministic in the future and in the past.

Finally, we impose that the $\mathscr{G}$ accessibility relation is a total preorder,

$$\begin{array}{c}\text{The accessibility relation } \mathscr{G} \text{ and its converse } \mathscr{G}^{-1} \\ \text{are reflexive, transitive and relatively total:} \\ \text{if } w_1, w_2 \in \mathscr{G}(w) \text{ then } w_1 \in \mathscr{G}(w_2) \text{ or } w_2 \in \mathscr{G}(w_1).\end{array} \qquad (\mathbf{SC_9})$$

This means that time is linear towards the past (by using $\mathscr{G}^{-1}$) and the future. One might object that at least future should be branching. For us, what is important is not the nature of time in the real world, but rather the perception that agents have of it. Thus, as time is linear here, each world believed to be possible by an agent can be

[5] Intuitively the elements of $\mathscr{P}_i(w)$ should also be "big" subsets of $\mathscr{B}_i(w)$: every $U \in \mathscr{P}_i(w)$ should contain more elements than its complement $\mathscr{B}_i(w) \setminus U$. But the language of modal logic is not expressive enough to account for this. The above constraint is therefore weaker, and there are neighborhoods satisfying our constraints gathering less than 50 % of the worlds (*cf.* [83]). However, this suffices to capture some interesting properties such as inconsistency of some emotions.

[6] This hypothesis does not permit to speak about counterfactual situations such as "if I had done α then φ would be true", but this is not problematic as long as we are not concerned with such hypothetic reasonings.

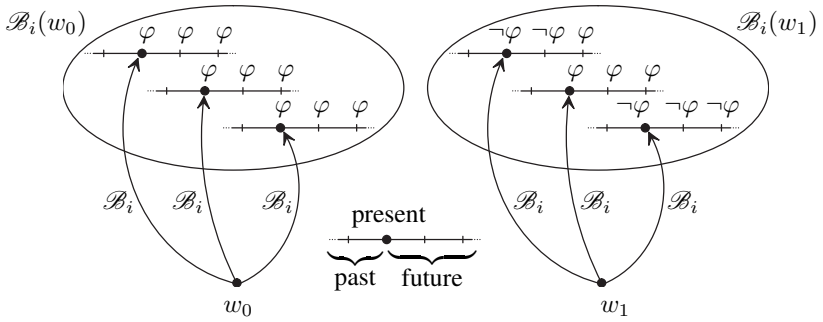

Figure 2: In world w_0, agent i *believes* that henceforth φ is true; in world w_1, for agent i there are three different possible histories: from top to bottom, the one where φ is currently false but it will occur in the future, the one where φ is henceforth true, and the one where φ is henceforth false.

identified with a history, that is a linear sequence of time points, and the diversity of futures is represented through different histories that are possible for the agent at the same world (*cf*. Figure 2). In other words, even if time is linear, several futures are possible for the agent, and we therefore have a *subjective version of branching-time*.

Moreover, we impose some constraints involving two or more accessibility relation types. In particular, we suppose that agents are aware of their probabilities and desirabilities, that is, the agents' beliefs about their own subjective probabilities and their desirabilities are correct and complete. We thus impose that, for every $i \in AGT$:

$$\text{if } w' \in \mathscr{B}_i(w) \text{ then } \mathscr{P}_i(w) = \mathscr{P}_i(w') \qquad \textbf{(SC}_{\mathbf{10}}\textbf{)}$$

$$\text{if } w' \in \mathscr{B}_i(w) \text{ then } \mathscr{D}_i(w) = \mathscr{D}_i(w') \qquad \textbf{(SC}_{\mathbf{11}}\textbf{)}$$

Concerning the relation between belief and action, we suppose that actions are public, in the sense that their occurrence is correctly and completely perceived by all agents. For every $i \in AGT$:

$$\text{if } w' \in \mathscr{B}_i(w) \text{ then } (\mathscr{A}_{j:\alpha})^{-1}(w) = \emptyset \text{ iff } (\mathscr{A}_{j:\alpha})^{-1}(w') = \emptyset \qquad \textbf{(SC}_{\mathbf{12}}\textbf{)}$$

We also impose that agents do not forget their previous alternatives ("no forgetting", alias "perfect recall" [30]). This relies in particular on the preceding hypothesis that actions are public, *viz*. that they are perceived correctly and completely by every agent. Thus, for every agent $i, j \in AGT$:

$$\text{if } (\mathscr{B}_i \circ \mathscr{A}_{j:\alpha})(w) \neq \emptyset \text{ then } (\mathscr{A}_{j:\alpha} \circ \mathscr{B}_i)(w) \subseteq (\mathscr{B}_i \circ \mathscr{A}_{j:\alpha})(w) \qquad \textbf{(SC}_{\mathbf{13}}\textbf{)}$$

In particular, it is true when i and j are the same agent. In terms of Figure 2, the agent's belief state after some action was performed at w_0, is a subset of the agent's belief state at w_0 that has been 'progressed' [70] in order to take into account the action occurrence.

Action and time are closely related. In particular, we impose that the future of every world w contains the worlds resulting from the performance of actions in w:

$$\mathscr{G} \supseteq \mathscr{A}_{i:\alpha} \text{ for each } i{:}\alpha \in AGT \times EVT \qquad (\mathbf{SC_{14}})$$

In words, the worlds resulting from the performance of actions in w are necessarily worlds in the future. But the converse is not necessarily true: every world in the future is not necessarily accessible by some action $i{:}\alpha$ in one step: such a hypothesis would be too strong.

For the sake of simplicity, we make the hypothesis that preferences are stable: what is desirable for an agent persists.[7]

$$\text{if } w\mathscr{G}w' \text{ then } \mathscr{D}_i(w) = \mathscr{D}_i(w') \qquad (\mathbf{SC_{15}})$$

This allows to disregard the influence of emotions on desirability. We are aware that our constraint is too strong in the general case, but it is quite realistic for rather short time intervals like a small dialog.

We make the same hypothesis for (social, legal, moral...) obligations, norms, standards... that hold for the agents:

$$\text{if } w\mathscr{G}w' \text{ then } \mathscr{I}(w) = \mathscr{I}(w'). \qquad (\mathbf{SC_{16}})$$

As we do not deal with the dynamics of ideals, it is quite reasonable to consider that ideals are stable, at least for a given time interval.

We call *EL frames* the set of frames satisfying constraints ($\mathbf{SC_1}$)–($\mathbf{SC_{16}}$).

3.2.3 EL models and validity

A model $\mathcal{M}$ is a couple $\langle \mathcal{F}, \mathscr{V} \rangle$ where:

- $\mathcal{F}$ is an EL frame;
- $\mathscr{V} : W \longrightarrow ATM$ associates each world w with the set $\mathscr{V}_w$ of atomic propositions true in w.

Given a model $\mathcal{M} = \langle \mathcal{F}, \mathscr{V} \rangle$ where $\mathcal{F} = \langle W, \mathscr{B}, \mathscr{P}, \mathscr{D}, \mathscr{I}, \mathscr{A}, \mathscr{G} \rangle$, we recursively define truth of a formula φ at a world w, noted $\mathcal{M}, w \models \varphi$ as follows:

- $\mathcal{M}, w \not\models \bot$;
- $\mathcal{M}, w \models p$ iff $p \in \mathscr{V}_w$;
- $\mathcal{M}, w \models \neg\varphi$ iff $\mathcal{M}, w \not\models \varphi$;
- $\mathcal{M}, w \models \varphi \vee \psi$ iff $\mathcal{M}, w \models \varphi$ or $\mathcal{M}, w \models \psi$;
- $\mathcal{M}, w \models Bel_i\,\varphi$ iff $\mathcal{M}, w' \models \varphi$ for every $w' \in \mathscr{B}_i(w)$;

[7] This allows concise statements and proofs of theorems (which else would have required the explicit statement of the relevant persistence hypotheses). In recent work we have relaxed this constraint in order to model emotion-focused coping strategies [4].

- $\mathcal{M}, w \models Prob_i\,\varphi$ iff there exists $U \in \mathscr{P}_i(w)$ such that for every $w' \in U$, $\mathcal{M}, w' \models \varphi$;
- $\mathcal{M}, w \models Des_i\,\varphi$ iff $\mathcal{M}, w' \models \varphi$ for every $w' \in \mathscr{D}_i(w)$;
- $\mathcal{M}, w \models Idl\,\varphi$ iff $\mathcal{M}, w' \models \varphi$ for every $w' \in \mathscr{I}(w)$;
- $\mathcal{M}, w \models After_{i:\alpha}\,\varphi$ iff $\mathcal{M}, w' \models \varphi$ for every $w' \in \mathscr{A}_{i:\alpha}(w)$;
- $\mathcal{M}, w \models Before_{i:\alpha}\,\varphi$ iff $\mathcal{M}, w' \models \varphi$ for every w' such that $w \in \mathscr{A}_{ia}(w')$;
- $\mathcal{M}, w \models G\varphi$ iff $\mathcal{M}, w' \models \varphi$ for every $w' \in \mathscr{G}(w)$;
- $\mathcal{M}, w \models H\varphi$ iff $\mathcal{M}, w' \models \varphi$ for every w' such that $w \in \mathscr{G}(w')$.

Validity of a formula φ in the class of all Kripke models obeying our semantic constraints is defined as usual. Thus, φ is true in model $\mathcal{M}$ if and only if $\mathcal{M}, w \models \varphi$ for every w in $\mathcal{M}$. φ is EL valid (noted $\models_{EL} \varphi$) if and only if φ is true in every EL model $\mathcal{M}$. φ is satisfiable if and only if $\not\models_{EL} \neg\varphi$. φ is a *logical consequence* of a set of (global) hypotheses Γ if and only if for every EL model $\mathcal{M}$, if all hypotheses of Γ are true in $\mathcal{M}$ then φ is true in $\mathcal{M}$.

3.3 Axiomatics

We now introduce a set of axioms that our modal operators have to satisfy. All our modal operators except $Prob_i$ are going to be normal modal operators, whose definition we recall first.

3.3.1 Normal operators

$\Box$ is a normal operator iff the axiom (K-$\Box$) and the necessitation rule (RN-$\Box$) hold for $\Box$.

$$\Box(\varphi \to \psi) \to (\Box\varphi \to \Box\psi) \qquad \text{(K-}\Box\text{)}$$

$$\frac{\varphi}{\Box\varphi} \qquad \text{(RN-}\Box\text{)}$$

In any normal modal logic, the semantics validates the following principles (which are used in some proofs in the appendix):

$$\frac{\varphi \to \psi}{\Box\varphi \to \Box\psi} \qquad \text{(RM-}\Box\text{)}$$

$$(\Box\varphi \wedge \Box\psi) \to \Box(\varphi \wedge \psi) \qquad \text{(C-}\Box\text{)}$$

The dual of $\Box$ is noted $\Diamond$ and obeys the following principle:

$$(\Box\varphi \wedge \Diamond\psi) \to \Diamond(\varphi \wedge \psi) \qquad (1)$$

and the following inference rule [18, Theorem 4.4, p. 116]:

$$\frac{\varphi \to \psi}{\Diamond\varphi \to \Diamond\psi} \qquad \text{(RM-}\Diamond\text{)}$$

More details on the formal properties of normal modal logics can be found in [18, Chapter 4].

3.3.2 Action

$After_\alpha$ and $Before_\alpha$ have the standard tense logic $\mathbf{K}_t$ in a linear time version: a normal modal logic **K** extended with the following axioms (*cf.* [16] for more details):

$$Happens_\alpha\,\varphi \to After_\beta\,\varphi \qquad \text{(CD-HA)}$$
$$Done_\alpha\,\varphi \to Before_\beta\,\varphi \qquad \text{(CD-DB)}$$
$$\varphi \to After_\alpha\,Done_\alpha\,\varphi \qquad \text{(CONV-AD)}$$
$$\varphi \to Before_\alpha\,Happens_\alpha\,\varphi \qquad \text{(CONV-BH)}$$

(CD-HA) and (CD-DB) are the axioms of common determinism. For example (CD-HA) means that if an action α is about to happen after which φ, then after any other action β, φ will be true, and similarly in the past for (CD-DB). This entails that actions take one time step, and are deterministic in the future and in the past (one can see that when α is β). The conversion axioms (CONV-AD) and (CONV-BH) link past and future.

Remember that $i{:}\alpha$ reads "agent i does action α".

We highlight here that what we call action is assumed to be intentional, that is the agent always intend to perform actions that he is about to perform. This is the difference between actions and events. Thus if an agents does something unintentionally (like sneezing) it is an event, and it can only trigger event-based emotions. This corresponds to Lazarus' control appraisal variable imposing that one can only reproach something to someone if this person had control over what she did, and to the the concept of responsibility in OCC theory [62, p. 54].

3.3.3 Belief

The operators Bel_i have the standard logic **KD45** (*cf.* [18] or [42] for more details). The corresponding axioms are those of normal modal logics plus the following ones:

$$Bel_i\,\varphi \to \neg Bel_i\,\neg\varphi \qquad \text{(D-}Bel_i\text{)}$$
$$Bel_i\,\varphi \to Bel_i\,Bel_i\,\varphi \qquad \text{(4-}Bel_i\text{)}$$
$$\neg Bel_i\,\varphi \to Bel_i\,\neg Bel_i\,\varphi \qquad \text{(5-}Bel_i\text{)}$$

Thereby an agent's beliefs are consistent (D-Bel_i), and an agent is aware of his beliefs (4-Bel_i) and disbeliefs (5-Bel_i).

3.3.4 Time

The operators G and H have the linear tense logic **S4.3**$_t$ (*cf.* [16]) which is a normal modal logic **K** for each operator plus the following axioms:

$$G\varphi \to \varphi \qquad \text{(T-}G\text{)}$$
$$(F\varphi \wedge F\psi) \to (F(\varphi \wedge F\psi) \vee F(\psi \wedge F\varphi)) \qquad \text{(3-}F\text{)}$$
$$G\varphi \to GG\varphi \qquad \text{(4-}G\text{)}$$
$$H\varphi \to \varphi \qquad \text{(T-}H\text{)}$$
$$(P\varphi \wedge P\psi) \to (P(\varphi \wedge P\psi) \vee P(\psi \wedge P\varphi)) \qquad \text{(3-}P\text{)}$$
$$H\varphi \to HH\varphi \qquad \text{(4-}H\text{)}$$
$$\varphi \to GP\varphi \qquad \text{(CONV-GP)}$$
$$\varphi \to HF\varphi \qquad \text{(CONV-HF)}$$

(T-G) and (T-H) mean that both future and past include the present.

(3-F) and (3-P) indicate that if two formulas are true at two instants in the future (resp. in the past) then one is necessarily true before the other. This entails that time is linear in the future and in the past (*cf.* Figure 2).

(CONV-GP) and (CONV-HF) are the conversion axioms. They axiomatize that the accessibility relation for G is the converse of that for H.

3.3.5 Probability

The notion of subjective probability measure is captured here semantically by the fact that probable worlds belong to the set of believed worlds. This approach is based on neighborhood functions (as opposed to probability distributions).

The logic of $Prob$ is weaker than the logic of belief. In particular, the formula $(Prob_i\,\varphi \wedge Prob_i\,\psi) \to Prob_i\,(\varphi \wedge \psi)$ is not valid, and this is enough to make it a non-normal modal logic in the sense of [18, Theorem 4.3].

The semantical conditions validate the following principles:

$$\frac{\varphi \to \psi}{Prob_i\,\varphi \to Prob_i\,\psi} \qquad \text{(RM-}Prob_i\text{)}$$
$$\frac{\varphi}{Prob_i\,\varphi} \qquad \text{(RN-}Prob_i\text{)}$$
$$Prob_i\,\varphi \to \neg Prob_i\,\neg\varphi \qquad \text{(D-}Prob_i\text{)}$$

3.3.6 Desirability

The logic of desirability is standard deontic logic (SDL) [18] and is also expressed in terms of ideal worlds: the logic associated with the operators Des_i is **KD**, *viz.* the normal modal logic **K** plus the following axiom:

$$Des_i\,\varphi \to \neg Des_i\,\neg\varphi \qquad \text{(D-}Des_i\text{)}$$

which makes desirabilities consistent.

It has been argued that in principle (*e.g.* [17] and also [45]), desirability is closed neither under implication nor under conjunction: It may be desirable for me to marry Ann and it may be desirable for me to marry Beth, but this does not imply that it is desirable for me to be a bigamist. Though, for the sake of simplicity, our Des_i operators are normal and hence closed under both conjunction and implication.

3.3.7 Ideals

Just as for desirability, the logic of ideality is standard deontic logic SDL, *viz.* the normal modal logic **K** plus the following axiom:

$$Idl\,\varphi \rightarrow \neg Idl\,\neg\varphi \qquad \text{(D-}Idl_i\text{)}$$

which makes ideals consistent.

3.3.8 Mix axioms

The interdependencies between some modal operators are captured by the following axioms. First, the following introspection axioms express that the agents are aware of their probabilities and desirabilies:

$$Prob_i\,\varphi \rightarrow Bel_i\,Prob_i\,\varphi \qquad \text{(4-MIX1)}$$
$$\neg Prob_i\,\varphi \rightarrow Bel_i\,\neg Prob_i\,\varphi \qquad \text{(5-MIX1)}$$
$$Des_i\,\varphi \rightarrow Bel_i\,Des_i\,\varphi \qquad \text{(4-MIX2)}$$
$$\neg Des_i\,\varphi \rightarrow Bel_i\,\neg Des_i\,\varphi \qquad \text{(5-MIX2)}$$

From these axioms plus (D-Bel_i), we can easily prove their converse. For example, we deduce the converse of (4-MIX1) from $Bel_i\,Prob_i\,\varphi \rightarrow \neg Bel_i\,\neg\,Prob_i\,\varphi$ by (D-Bel_i), and $\neg Bel_i\,\neg Prob_i\,\varphi \rightarrow Prob_i\,\varphi$ by (5-MIX1). We therefore have the equivalences $Prob_i\,\varphi \leftrightarrow Bel_i\,Prob_i\,\varphi$ and $\neg Prob_i\,\varphi \leftrightarrow Bel_i\,\neg Prob_i\,\varphi$.

Then the following axioms express that actions are public:

$$Done_\alpha\,\top \rightarrow Bel_i\,Done_\alpha\,\top \qquad \text{(4-MIX3)}$$
$$\neg Done_\alpha\,\top \rightarrow Bel_i\,\neg Done_\alpha\,\top \qquad \text{(5-MIX3)}$$

From these axioms plus (D-Bel_i) we can easily prove their converse, and we thus have the equivalences $Done_\alpha\,\top \leftrightarrow Bel_i\,Done_\alpha\,\top$ and $\neg Done_\alpha\,\top \leftrightarrow Bel_i\,\neg Done_\alpha\,\top$.

We axiomatize the inclusion of elements of neighborhoods in epistemic states through the following axiom:

$$(Bel_i\,\varphi \wedge Prob_i\,\psi) \rightarrow Prob_i\,(\varphi \wedge \psi) \qquad \text{(C-MIX)}$$

which allows to derive the following theorems:

$$Bel_i\,\varphi \rightarrow Prob_i\,\varphi \qquad (2)$$
$$Prob_i\,\varphi \rightarrow \neg Bel_i\,\neg\varphi \qquad (3)$$

Time and action are linked: if φ is always true in the future then φ will be true after every action performance. Similarly, if φ was always true in the past, then φ was true before every performance of an action. So:

$$G\varphi \rightarrow After_{\alpha}\,\varphi \qquad \text{(GA-MIX)}$$
$$H\varphi \rightarrow Before_{\alpha}\,\varphi \qquad \text{(HB-MIX)}$$

Finally, desirability persists, i.e. it is preserved through time.

$$Des_i\,\varphi \rightarrow G\,Des_i\,\varphi \qquad \text{(Pers-}Des_i\text{)}$$
$$\neg Des_i\,\varphi \rightarrow G\neg Des_i\,\varphi \qquad \text{(Pers-}\neg Des_i\text{)}$$

These two principles both entail the equivalences $Des_i\,\varphi \leftrightarrow G\,Des_i\,\varphi$ and $\neg Des_i\,\varphi \leftrightarrow G\neg Des_i\,\varphi$.

For the same reasons, ideals also persist:

$$Idl\,\varphi \rightarrow G\,Idl\,\varphi \qquad \text{(Pers-}Idl_i\text{)}$$
$$\neg Idl\,\varphi \rightarrow G\neg Idl\,\varphi \qquad \text{(Pers-}\neg Idl_i\text{)}$$

These two principles entail that we have an equivalence.

The "no forgetting" constraint linking actions and belief is captured by the following axiom:

$$(Bel_i\,After_{\alpha}\,\varphi \wedge \neg Bel_i\,After_{\alpha}\,\bot) \rightarrow After_{\alpha}\,Bel_i\,\varphi \qquad \text{(NF-}Bel_i\text{)}$$

This axiom expresses that the agents do not forget their previous alternatives, when the performance of the action is not surprising for them ($\neg Bel_i\,After_{\alpha}\,\bot$ reads "agent i does not believe that action α is inexecutable"). Otherwise, if $Bel_i\,After_{\alpha}\,\bot$ holds, then the agent has to revise his beliefs upon learning that α occurred. We do not go into this here, and refer the reader to [40].

In the next two sections we are going to put to work logic EL, and are going to express twenty from the twenty-two emotions of OCC theory. (We do not define the remaining emotions of love and hate because they would require a first order modal logic.) For each of these twenty emotions, we first give the informal definition of OCC theory, and then our definition in terms of logical formulas. In order to support the accuracy of our definitions, we show that they can account for the examples illustrating the emotions in [62]. Below, the quoted pages all refer to this book.

4 Event-based emotions

The event-based branch of OCC theory contains emotion types whose eliciting conditions depend on the evaluation of an event with respect to the agent's goals. *Desirability* is the central variable accounting for the impact that an event has on an agent's goals, namely how it helps or impedes their achievement.

In our formalism, an event is something that may occur without any agent intending it, and is thus different from an action (that is always intentional). According to

OCC theory an event can have several aspects, each of them possibly triggering a different emotion. In this paper we represent an emotion as an abbreviation of a complex formula. Moreover we assume that what Ortony *et al*. call the different aspects of an event can be considered as consequences of the primary event. For example the event of receiving a letter from a bailiff to inform you that you are going to inherit some money from a deceased relative has (at least) two aspects: the undesirable aspect is that your relative is dead, while the desirable aspect is that you get some money. We represent these two aspects as if they were two separate secondary events actually resulting from the primary event. While the same primary event can trigger opposite emotions (sadness that your relative died and joy of getting some money), we consider in our formalization that these emotions are attached to two different secondary events, but not to the primary event.

According to OCC theory, an event is desirable for an agent if its consequence φ is more beneficial (furthering his goals) than harmful (impeding some goals). As we said before (see Section 2.2), desirability depends on the agent's goals, but we do not want to enter into the details of this computation here, and assume that the agent's desirability values are given by the Des operators. We directly use this variable in the definitions of event-based emotions (*cf.* Section 3 for our modelling of desirability).

4.1 Well-being emotions

The emotion types in this group have eliciting conditions focused on the desirability for the self of an event. An agent feels joy (resp. distress) when he is pleased (resp. displeased) about a desirable (resp. undesirable) event.

$$Joy_i\,\varphi \stackrel{def}{=} Bel_i\,\varphi \wedge Des_i\,\varphi$$
$$Distress_i\,\varphi \stackrel{def}{=} Bel_i\,\varphi \wedge Des_i\,\neg\varphi$$

Consider an example situation from [62, p. 88] where a man i learns that he inherits of a small amount of money (*m*) from a remote and unknown relative that has died (*d*). This is expressed by the formula $Bel_i\,(m \wedge d)$. Then i feels **joy** because he focuses on the desirable event ($Des_i\,m$) and not on the undesirable event *d*. This man does not feel distress about his relative's death since he did not know the relative, his death is not undesirable for him ($\neg Des_i\,\neg d$). On the contrary, a man j (p. 89) who runs out of gas on the freeway ($Bel_j\,o$) feels **distress** because this is undesirable for him ($Des_j\,\neg o$).

4.2 Prospect-based emotions

The emotion types in this group have eliciting conditions focused on the desirability for self of an anticipated (uncertain) event, that is actively prospected. OCC uses a local intensity variable called *likelihood*, accounting for the expected probability of the event to occur. We model likelihood by the following abbreviation $Expect_i$.

Definition 1 $Expect_i\,\varphi \stackrel{def}{=} Prob_i\,\varphi \wedge \neg Bel_i\,\varphi$

$Expect_i\,\varphi$ reads "agent i expects φ to be true but envisages that it could be false". We can notice that if i expects something then he necessarily envisages it:

$$Expect_i\,\varphi \rightarrow \neg Bel_i\,\neg\varphi \tag{4}$$

From (D-$Prob_i$) we can easily prove the consistency of expectations:

$$Expect_i\,\varphi \rightarrow \neg Expect_i\,\neg\varphi \tag{5}$$

An agent feels hope (resp. fear) if he is "pleased (resp. displeased) about the **prospect** of a desirable (resp. undesirable) event". Note that the object of hope is not necessarily about the future: I might ignore whether my email has been delivered to the addressee, and hope it has been so.

$$Hope_i\,\varphi \stackrel{def}{=} Expect_i\,\varphi \wedge Des_i\,\varphi$$
$$Fear_i\,\varphi \stackrel{def}{=} Expect_i\,\varphi \wedge Des_i\,\neg\varphi$$

The agent feels fear-confirmed (resp. satisfaction) if he is "displeased (resp. pleased) about the **confirmation** of the prospect of an undesirable (resp. desirable) event". We use here our operator P (see Definition Def$_P$, just before Section 3.2) to represent what was true in the past.

$$Satisfaction_i\,\varphi \stackrel{def}{=} Bel_i\,PExpect_i\,\varphi \wedge Des_i\,\varphi \wedge Bel_i\,\varphi$$
$$FearConfirmed_i\,\varphi \stackrel{def}{=} Bel_i\,PExpect_i\,\varphi \wedge Des_i\,\neg\varphi \wedge Bel_i\,\varphi$$

Given our definitions of joy and distress $Satisfaction_i\,\varphi$ can be written more concisely $Bel_i\,PExpect_i\,\varphi \wedge Joy_i\,\varphi$, and $FearConfirmed_i\,\varphi$ can be written $Bel_i\,PExpect_i \wedge Distress_i\,\varphi$.

The agent feels relief (resp. disappointment) if he is "pleased (resp. displeased) about the **disconfirmation** of the prospect of an undesirable (resp. desirable) event".

$$Relief_i\,\varphi \stackrel{def}{=} Bel_i\,PExpect_i\,\neg\varphi \wedge Des_i\,\varphi \wedge Bel_i\,\varphi$$
$$Disappointment_i\,\varphi \stackrel{def}{=} Bel_i\,PExpect_i\,\neg\varphi \wedge Des_i\,\neg\varphi \wedge Bel_i\,\varphi$$

For example a woman w who applies for a job (p. 111) might feel **fear** if she expects not to be offered the job ($Expect_w$ *¬beHired*), or feel **hope** if she expects that she will be offered it ($Expect_w$ *beHired*). Then, if she hoped to get the job and finally gets it, she feels **satisfaction**; and if she does not get it, she feels **disappointment**. An employee e (p. 113) who expects to be fired ($Expect_e$ *beFired*) will feel **fear** if it is undesirable for him (Des_e *¬beFired*), but not if he already envisaged to quit this job since in this case we can suppose that this is not undesirable for him ($\neg Des_e$ *¬beFired*). In the first case he will feel **relief** when he is not fired (Bel_e *¬beFired*), and **fear-confirmed** when he is.

4.3 Fortunes-of-others emotions

The emotion types in this group have eliciting conditions focused on the presumed desirability for another agent. They use three local intensity variables: *desirability*

for other, *deservingness*, and *liking*. *Desirability for other* is the assessment by i of how much the event is desirable for the other one (j). *Deservingness* represents how much agent i believes that agent j deserved what occurred to him. (It often depends on *liking*, *viz.* i's attitude towards j, but we cannot account for this here because we do not consider attraction emotions.)

We thus have to model these variables. First, we can represent *desirability for other* by a belief about the other's desire: $Bel_i\, Des_j\, \varphi$ reads "agent i believes that φ is desirable for agent j". Second, we represent *liking* through non-logical axioms. For example, when John likes Mary this means that if John believes that it is desirable for Mary to be rich, then it is desirable for John that Mary be rich, or better: gets to know that she is rich. In formulas: $Bel_{john}\, Des_{mary}\, rich \rightarrow Des_{john}\, Bel_{mary}\, rich$). These axioms will be global hypotheses in deductions, in the sense that they are supposed to be known by all agents and to hold through time. Third, we simplify the concept of *deservingness* by assuming that agents want (have a goal) that any agent gets what he deserves. Then when an event (represented with the formula φ) occurs that is believed (by i) to be deserved by j, this event will be desirable for i (we write it $Des_i\, Bel_j\, \varphi$[8]), since it achieves his goal. So finally i can desire that j believes φ either because he believes that j desires φ and j is his friend, or because he believes that j desires $\neg\varphi$ and j is his enemy, or because he believes that j deserved φ.

There are two good-will (or empathetic) emotions: an agent feels happy for (resp. sorry for) another agent if he is pleased (resp. displeased) about an event presumed to be desirable (resp. undesirable) for this agent.

$$HappyFor_{i,j}\varphi \stackrel{def}{=} Bel_i\, \varphi \wedge Bel_i\, Des_j\, \varphi \wedge Des_i\, Bel_j\, \varphi$$
$$SorryFor_{i,j}\varphi \stackrel{def}{=} Bel_i\, \varphi \wedge Bel_i\, Des_j\, \neg\varphi \wedge Des_i\, \neg Bel_j\, \varphi$$

There are two ill-will emotions: an agent feels resentment (resp. gloating) towards another agent if he is displeased (resp. pleased) about an event presumed to be desirable (resp. undesirable) for this agent.

$$Resentment_{i,j}\varphi \stackrel{def}{=} Bel_i\, \varphi \wedge Bel_i\, Des_j\, \varphi \wedge Des_i\, \neg Bel_j\, \varphi$$
$$Gloating_{i,j}\varphi \stackrel{def}{=} Bel_i\, \varphi \wedge Bel_i\, Des_j\, \neg\varphi \wedge Des_i\, Bel_j\, \varphi$$

For example (p. 95) Fred feels **happy for** Mary when he learns that she wins a thousand dollars ($Bel_f\, w \wedge Bel_f\, Bel_m\, w$), because he believes this is desirable for her ($Bel_f\, Des_m\, w$), and she is his friend, *viz.* he *likes* her. As said above, we represent this notion of *liking* with non-logical global axioms representing one's interest in the well-being of one's friends. In this case, if Fred believes that it is desirable for Mary to win, then it is desirable for him that she gets to know that she won ($Bel_f\, Des_m\, w \rightarrow$

[8]The emotions types in this group result from the appraisal of an event **concerning another agent**. We represent the occurrence of this event φ to agent j by the formula $Bel_j\, \varphi$. Then the $Des_i\, Bel_j\, \varphi$ element of the definitions means that this event occurring to j is desirable for i, which is our way to distinguish between good-will and ill-will emotions.

$Des_f\ Bel_m\ w$).[9]

A man i (p. 95) can feel **sorry for** the victims v of a natural disaster ($Bel_i\ disaster \wedge Bel_i\ Bel_v\ disaster \wedge Bel_i\ Des_v\ \neg disaster$) without even knowing them, because he has an interest that people do not suffer undeservedly, so there it is undesirable for him that people suffer from this disaster ($Des_i\ \neg Bel_v\ disaster$).

An employee e (p. 99) can feel **resentment** towards a colleague c who received a large pay raise ($Bel_e\ pr,\ Bel_e\ Des_c\ pr$), what he believes to be desirable for this colleague ($Bel_c\ Des_e\ pr$), because he thinks this colleague is incompetent and thus does not deserve this raise. As we said above, we represent the deservingness (whatever the reason for this belief, here the incompetence) with a desirability concerning the occurrence of this event to the other agent (here $Des_e\ \neg Bel_c\ pr$) that is its consequence.

Finally, Nixon's political opponents (o) (p. 104) might have felt **gloating** about his departure from office ($Bel_o\ d \wedge Bel_o\ Bel_{nixon}\ d$), because they believed it to be undesirable for him ($Bel_o\ Des_{nixon}\ \neg d$) and they thought it was deserved (as above we identify deservingness with a desire, here: $Des_o\ Bel_{nixon}\ d$).

Remark 1 *Our formalization of liking leads to the following question: what if i believes that for some reasons j will never learn that φ? In many situations it is certainly odd to say that i is happy or sorry for another agent j about something that j will never know, and thus about what j will never be happy or sad about himself. OCC theory does textbfnot require that j should know about this event that is important to him, and we therefore have chosen to stay as close as possible to it. However, one might wish to sharpen the definitions of these four emotions by requiring that it must be at least probable for i that j learns about the event at some time point in the future. This can be implemented by adding the further conjunct $Prob_i\ F\ Bel_j\ \varphi$ to our definitions of the four fortunes-of-others emotions.*

5 Agent-based emotions

The agent-based branch of OCC theory contains emotion types whose eliciting conditions depend on the judgement of the praiseworthiness of an action, with respect to standards.

In our sense an action is something that is performed intentionally **(deliberately, purposely)** by an agent. It thus differs from an event. If an agent performs an action not purposely, like sneezing, we call this an event. This distinction allows to capture implicitly Lazarus' variable of attribution of responsibility that is needed for emotions like anger: an agent is always responsible for his actions.

An action is *praiseworthy* (resp. *blameworthy*) when it upholds (resp. violates) standards. The standards under concern are supposed to be internalized, i.e. the (evaluating) agent has adopted them. We express these internalized standards for agent i through the deontic operators Idl_i .

[9] Note that Fred may not be happy for Mary if she was not to learn about her gain in the future. However, even if she does not know yet that she won, he can feel happy for her just because he considers it probable that she will learn it at a future moment (without being sure of that). For example, Mary may have not seen the results yet, and Fred cannot be sure that she will not forget to check them.

5.1 Attribution emotions

The emotion types in this group have eliciting conditions focused on the approving of an agent's action. They use two local intensity variables. *Strength of unit* intervenes in self-agent emotions to represent the degree to which the agent identifies himself with the author of the action, allowing him to feel pride or shame when he is not directly the actor; for example one can be proud of his son succeeding in a difficult examination, or of his rugby team winning the championship; in this paper we only focus on emotions felt by the agent about his own actions, because this variable is too complex to be represented in our framework. *Expectation deviation* accounts for the degree to which the performed action differs from what is usually expected from the agent, according to his social role or category[10].

We express this notion of expectation with the formula $Prob_j\ After_{i:\alpha}\ \neg\varphi$ reading "j considers it probable that after i performs α, φ is false", *viz.* j expects i not to achieve φ as a result of his action, for example because it is difficult.[11] The deviation comes from the fact that after the execution of α, j believes that φ is nevertheless true, contrarily to what he expected.[12] This prevents the agent from feeling attribution emotions too often. Indeed, we often respect the law without being proud, and we often violate standards without being ashamed. Therefore we consider that the standards have to be internalized and accepted by the agent as belonging to his values. This allows an agent to feel no emotion, even concerning an (un)ideal action, when this is not important for him. For example someone who likes to wear strange (unideal) clothes would not feel ashamed about this if it is what he desires to wear, but would feel so if he was forced to wear such clothes.

Finally, we do not impose that the ideal was conscious at the moment of the action. For example one can feel shame about having performed an action when one realizes that it was blameworthy, even if one ignored that at the time when the action was performed. Ideally, we should not impose it either for probability, but the $Prob_i$ operators are intrinsically epistemic (*viz.* semantically, probable worlds are a subset of possible worlds compatible with the agent's beliefs); so technically it is difficult to do so.

In the sequel, $Emotion_i(i{:}\alpha, \varphi)$ (resp. $Emotion_{i,j}(j{:}\alpha, \varphi)$) abbreviates $Emotion_i(Done_{i:\alpha}\ \top, \varphi)$ (resp. $Emotion_{i,j}(Done_{j:\alpha}\ \top, \varphi)$) where $Emotion$ is the name of an emotion.

Remark 2 *These emotions are about an action α that the agent believes to have influenced the proposition φ: the agent believes that "if he had not performed action α, φ would probably be false now". Though, our language is not expressive enough to represent this counterfactual reasoning, so we make the hypothesis that the agent i believes that α and φ are linked in this way. The following emotions do make sense only when this is the case.*

[10] In self-agent emotions, the agent refers to his stereotyped representation of himself.

[11] In the following, whatever the cause of the unexpectedness is (for example difficulty), we only formalize the consequence (the unexpectedness itself) with the above formula.

[12] What is unexpected is not only the performance of the action but also its result φ; actually, when the result is not important, φ is $\top$ and then it is the very performance of the action that is unexpected. For example it is unexpected from a wise child to steal something in a shop, whatever the result of his action is (did he succeed or not), so we write: $Prob_i\ After_{child:steal}\ \bot$.

Self-agent emotions: an agent feels pride (resp. shame) if he is approving (resp. disapproving) of his own praiseworthy (resp. blameworthy) action.

$$\begin{aligned}
&\mathit{Pride}_i\,(i{:}\alpha, \varphi) \stackrel{def}{=} \\
&\qquad \mathit{Bel}_i\,\mathit{Done}_{i:\alpha}\,(\mathit{Idl}_i\,\mathit{Happens}_{i:\alpha}\,\varphi \wedge \mathit{Prob}_i\,\mathit{After}_{i:\alpha}\,\neg\varphi) \wedge \mathit{Bel}_i\,\varphi \\
&\mathit{Shame}_i\,(i{:}\alpha, \varphi) \stackrel{def}{=} \\
&\qquad \mathit{Bel}_i\,\mathit{Done}_{i:\alpha}\,(\mathit{Idl}_i\,\neg\mathit{Happens}_{i:\alpha}\,\varphi \wedge \mathit{Prob}_i\,\mathit{After}_{i:\alpha}\,\neg\varphi) \wedge \mathit{Bel}_i\,\varphi
\end{aligned}$$

Emotions involving another agent:[13] an agent feels admiration (resp. reproach) towards another agent if he is approving (resp. disapproving) of this agent's praiseworthy (resp. blameworthy) action.

$$\begin{aligned}
&\mathit{Admiration}_{i,j}(j{:}\alpha, \varphi) \stackrel{def}{=} \\
&\qquad \mathit{Bel}_i\,\mathit{Done}_{j:\alpha}\,(\mathit{Idl}_i\,\mathit{Happens}_{j:\alpha}\,\varphi \wedge \mathit{Prob}_i\,\mathit{After}_{j:\alpha}\,\neg\varphi) \wedge \mathit{Bel}_i\,\varphi \\
&\mathit{Reproach}_{i,j}(j{:}\alpha, \varphi) \stackrel{def}{=} \\
&\qquad \mathit{Bel}_i\,\mathit{Done}_{j:\alpha}\,(\mathit{Idl}_i\,\neg\mathit{Happens}_{j:\alpha}\,\varphi \wedge \mathit{Prob}_i\,\mathit{After}_{j:\alpha}\,\neg\varphi) \wedge \mathit{Bel}_i\,\varphi
\end{aligned}$$

For example, a woman m feels **pride** (p. 137) of having saved the life of a drowning child because she performed the action α (to jump into the water to try to save him) with the successful result s (the child is safe): $\mathit{Bel}_m\,\mathit{Done}_{m:\alpha}\,\top \wedge \mathit{Bel}_m\,s$.[14] Moreover she now believes that before the action, it was ideal to save the child and she internalized this ideal ($\mathit{Idl}_m\,\mathit{Happens}_{m:\alpha}\,\top$), but she had not much chances to succeed:[15] $\mathit{Prob}_m\,\mathit{After}_{m:\alpha}\,\neg s$.

A rich elegant lady l (p. 142) would feel **shame** when caught while stealing clothes in an exclusive boutique ($\mathit{Shame}_l\,(\alpha, \top)$, where α is the action to steal, because she has performed an action that was unideal for her[16] ($\mathit{Idl}_l\,\neg\mathit{Happens}_{l:\alpha}\,\top$) and improbable to be performed by her ($\mathit{Prob}_l\,\mathit{After}_{l:\alpha}\,\bot$) due to her social role. The result of the action is $\top$ here because this emotion does not depend on the success or failure of the action but on its very performance.

A physicist p's colleagues c (p. 145) feel **admiration** towards him for his Nobel-prize-winning work ($\mathit{Bel}_c\,\mathit{Done}_{p:\alpha}\,\top \wedge \mathit{Bel}_c\,w$, where α is the action of conducting experiments, with the result w of obtaining Nobel-prize-deserving findings) because they internalized this result as ideal.[17] ($\mathit{Idl}_c\,\mathit{Happens}_{p:\alpha}\,w$) and difficult thus unex-

[13] When $i = j$, these emotions correspond to the self-agent emotions (*cf.* Theorem 6).

[14] Actually, she also believes that she influenced this result by her action, *viz.* she believes that if she had not jumped into the water the child could have drowned; as we said it before, we cannot express this causal link in our language, so our account is incomplete in that respect.

[15] Thus, she would not feel pride after saving the child if she believes it was easy for her.

[16] Actually actions do not obligatorily follow moral values. The lady may have been driven by the desire to possess the object, violating her ideals. But this example seems to be a borderline case, since she could have bought the object instead.

[17] Here, what is ideal is not only the execution of the action but its execution with this result. Similarly, in the case of negative emotions, what is unideal is not the happening of the action, but its happening with a given result: $\mathit{Idl}_i\,\neg\mathit{Happens}_{i:\alpha}\,\varphi$. This is compatible with the fact that the action itself could be ideal: $\mathit{Idl}_i\,\mathit{Happens}_{i:\alpha}\,\top$. For example, it is ideal to participate, but unideal to lose when you are expected to win.

pected ($Prob_c\ After_{p:\alpha}\ \neg w$). As we said above, the difficulty of an action is one possible reason for its result to be unexpected. Here to simplify we do not formalize the very notion of difficulty but only its consequence, *viz.* the unexpectedness of the result, which is what we are interested in when we define attribution emotions.

A man i may feel **reproach** towards a driver j (p. 145) who drives without a valid license ($Bel_i\ Done_{j:\delta}\ \top$, where δ is the action to drive without a valid license), because it is forbidden and he considers this obligation to be important ($Idl_i\ \neg Happens_{j:\delta}\ \top$) and unexpected from a driver ($Prob_i\ After_{j:\delta}\ \bot$).

5.2 Compound emotions

These emotions occur when the agent appraises both the consequences of the event and its agency. They are thus the result of a combination of attribution emotions about an action α with result φ, and well-being emotions about this result φ.

$$Gratification_i\,(i{:}\alpha, \varphi) \stackrel{def}{=} Pride_i\,(i{:}\alpha, \varphi) \wedge Joy_i\,\varphi$$

$$Remorse_i\,(i{:}\alpha, \varphi) \stackrel{def}{=} Shame_i\,(i{:}\alpha, \varphi) \wedge Distress_i\,\varphi$$

$$Gratitude_{i,j}(j{:}\alpha, \varphi) \stackrel{def}{=} Admiration_{i,j}(j{:}\alpha, \varphi) \wedge Joy_i\,\varphi$$

$$Anger_{i,j}(j{:}\alpha, \varphi) \stackrel{def}{=} Reproach_{i,j}(j{:}\alpha, \varphi) \wedge Distress_i\,\varphi$$

For example, a woman i may feel **gratitude** (p. 148) towards the stranger j who saved her child from drowning ($Bel_i\ Done_{j:\alpha}\ \top \wedge Bel_i\ s$, where $j{:}\alpha$ is j's action to jump in the water, and s is the result: her child is safe). Indeed, i feels admiration towards j because of j's ideal but difficult (*viz.* before it, $Prob_i\ After_{j:\alpha}\ \neg s$ held) action. Moreover the result of j's action ($Bel_i\ s$) is desirable for i ($Des_i\ s$), so i also feels joy about it ($Joy_i\ s$).

Similarly, a woman w (p. 148) may feel **anger** towards her husband h who forgets to buy the groceries ($Bel_w\ Done_{h:\alpha}\ \top$, where α is his action to go shopping, and $Bel_w\ \neg g$, where g reads "there are groceries for dinner"), because w reproaches this unideal result to h (it was not the expected result of the action: $Prob_w\ After_{h:\alpha}\ g$), and she is also sad about it ($Distress_w\ \neg g$) because she desired to eat vegetables ($Des_w\ g$).

The physicist p may feel **gratification** about winning the Nobel prize because he performed a successful execution of action α (performing experiments), achieving the ideal result n (he receives the Nobel prize), and thus feels pride; and this result is not only socially ideal but also desirable for him[18] ($Des_p\ n$), so pride combines with joy.

Finally, a spy may feel **remorse** (p. 148) about having betrayed his country (action ω) if he moreover caused undesirable damages (result d): $Shame_{spy}\,(\omega, d) \wedge Distress_{spy}\,d$.

6 Formal properties

In the previous section we started from OCC theory, extracted its key concepts, and casted them into logical definitions of twenty emotions. The first benefit of our work

[18]This is not always true. For example, a child may personally desire not to go to school, while it is ideal to go.

is to disambiguate these definitions, that might be debatable when expressed in natural language. For example, we had to decide between two different options in the case of fortunes-of-others emotions, depending on whether we accepted that one can feel happy about some good news for somebody else even if we believe that this person will never learn the good news. Furthermore, the use of logic enables to reason about the formalized concepts and to derive properties. In contrast, properties of emotions are always debatable when defined informally, and many debates have occurred as research progressed.

In this section we expose some theorems following from our definitions, mainly concerning the causal and temporal links that emotions have with each other. These theorems are consistent with OCC theory; sometimes they even go beyond it, but they always remain intuitive. Moreover, what is interesting is that the formal proofs of these theorems make that they are not debatable on their own once one has accepted the principles of the logic. This shows again the advantages of formal reasoning about emotions. The reader who is interested in the proofs of these theorems is referred to the appendix.

6.1 Prospect-based emotions and their confirmation

If an agent remembers that at a moment in the past he was feeling a prospect-based emotion about φ, and if he now knows whether φ is true or false, then it follows by the laws of our logic that he feels the corresponding confirmation emotion.

Theorem 1 (Temporal link from prospect to confirmation)

$$\vdash (Bel_i\, PHope_i\, \varphi \wedge (Bel_i\, \varphi \vee Bel_i\, \neg\varphi)) \rightarrow Satisfaction_i\, \varphi \vee Disappointment_i\, \neg\varphi \quad \text{(a)}$$

$$\vdash (Bel_i\, PFear_i\, \varphi \wedge (Bel_i\, \varphi \vee Bel_i\, \neg\varphi)) \rightarrow Relief_i\, \varphi \vee FearConfirmed_i\, \neg\varphi \quad \text{(b)}$$

Moreover, we can prove that an agent cannot feel simultaneously two emotions concerning the confirmation and the disconfirmation of the same expectation.

Theorem 2 (Inconsistency between confirmation and disconfirmation)

$$\vdash \neg(Satisfaction_i\, \varphi \wedge Disappointment_i\, \neg\varphi) \quad \text{(a)}$$

$$\vdash \neg(FearConfirmed_i\, \varphi \wedge Relief_i\, \neg\varphi) \quad \text{(b)}$$

The proof follows from the rationality axiom for belief.

Please note that on the contrary, we cannot prove inconsistencies between relief and satisfaction, or between fear-confirmed and disappointment. This is because $Bel_i\, PExpect_i\, \neg\varphi$ and $Bel_i\, PExpect_i\, \varphi$ are consistent, *viz.* the agent may have expected φ at one moment in the past and $\neg\varphi$ at another moment.[19] We can only prove that these two expectations

[19] Thus, our current definitions of confirmation and disconfirmation emotions may not be precise enough to entail this intuitive inconsistency. Actually in linear temporal logic with $Until$ and $Since$ operators, we could write for example $Relief_i\, \varphi \stackrel{def}{=} Bel_i\, P(\neg Expect_i\, \neg\varphi\, Since\, Expect_i\, \varphi) \wedge Des_i\, \varphi \wedge Bel_i\, \varphi$.

$Expect_i\ \neg\varphi$ and $Bel_i\ PExpect_i\ \varphi$ cannot occur at the same time. (This is the theorem (5) of Section 4.2).

We can prove that the positive confirmation emotions imply joy, and that the negative confirmation emotions imply distress. This is intuitive, and in agreement with Ortony *et al.*'s definitions.

Theorem 3 (Link between confirmation and well-being emotions)

$$\vdash Satisfaction_i\ \varphi \rightarrow Joy_i\ \varphi \quad \text{(a)}$$

$$\vdash FearConfirmed_i\ \varphi \rightarrow Distress_i\ \varphi \quad \text{(b)}$$

$$\vdash Relief_i\ \varphi \rightarrow Joy_i\ \varphi \quad \text{(c)}$$

$$\vdash Disappointment_i\ \varphi \rightarrow Distress_i\ \varphi \quad \text{(d)}$$

6.2 Fortunes-of-others emotions

In this paragraph we will ground on reinforced definitions of fortunes-of-others emotions, that we denote them by $Emotion'_{i,j}\varphi$ where $Emotion'$ ranges over the four fortunes-of-other emotions in $\{HappyFor, SorryFor,\ Resentment, Gloating\}$. These reinforced definitions are obtained from our definitions by adding the further conjunct $Prob_i\ FBel_j\ \varphi$ to them, as we suggested in Remark 1. For instance $HappyFor'_{i,j}\varphi \stackrel{def}{=} HappyFor_{i,j}\varphi \wedge Prob_i\ FBel_j\ \varphi$.

We can prove that if the agent i feels a fortune-of-other emotion towards another agent j about φ, then it is at least probable for i that j is going to feel the corresponding well-being emotion about φ at some moment in the future.

This leads us to believe that OCC definitions of these emotions may be to vague, since they do not allow to deduce these properties while they are quite intuitive.

Theorem 4 (From fortune-of-other emotion to image of other)

$$\vdash HappyFor'_{i,j}\varphi \rightarrow Prob_i\ FJoy_j\ \varphi \quad \text{(a)}$$

$$\vdash SorryFor'_{i,j}\varphi \rightarrow Prob_i\ FDistress_j\ \varphi \quad \text{(b)}$$

$$\vdash Resentment'_{i,j}\varphi \rightarrow Prob_i\ FJoy_j\ \varphi \quad \text{(c)}$$

$$\vdash Gloating'_{i,j}\varphi \rightarrow Prob_i\ FDistress_j\ \varphi \quad \text{(d)}$$

If an agent i feels a fortune-of-other emotion towards another agent about φ, and i is not sure that j will learn about the event φ, then i feels a corresponding prospect-based emotion about j believing φ.

Theorem 5 (Consequences of fortunes-of-others emotions)

$$\vdash (HappyFor'_{i,j}\varphi \wedge \neg Bel_i\ FBel_j\ \varphi) \rightarrow Hope_i\ FBel_j\ \varphi \quad \text{(a)}$$

$$\vdash (SorryFor'_{i,j}\varphi \wedge \neg Bel_i\ FBel_j\ \varphi) \rightarrow Fear_i\ FBel_j\ \varphi \quad \text{(b)}$$

$$\vdash (Resentment'_{i,j}\varphi \wedge \neg Bel_i\ FBel_j\ \varphi) \rightarrow Fear_i\ FBel_j\ \varphi \quad \text{(c)}$$

$$\vdash (Gloating'_{i,j}\varphi \wedge \neg Bel_i\ FBel_j\ \varphi) \rightarrow Hope_i\ FBel_j\ \varphi \quad \text{(d)}$$

6.3 Links between self-agent and other-agent attribution emotions

We can prove that an other-agent emotion towards oneself is equivalent to the corresponding self-agent emotion. This is rather intuitive, all the more Ortony *et al*. introduce the term *self-reproach* for shame.

Theorem 6 (Other-agent emotions towards oneself)

$$\vdash \mathit{Admiration}_{i,i}(i{:}\alpha, \varphi) \leftrightarrow \mathit{Pride}_i\,(i{:}\alpha, \varphi) \quad \text{(a)}$$

$$\vdash \mathit{Reproach}_{i,i}(i{:}\alpha, \varphi) \leftrightarrow \mathit{Shame}_i\,(i{:}\alpha, \varphi) \quad \text{(b)}$$

We can prove that if another agent j feels an attribution emotion towards an agent i about a given action with a given result, then the agent i does not inevitably feel the corresponding self-agent attribution emotion. That is, one can admire you about a given action while you are not proud about it.

Theorem 7 (Other-agent emotion does not force self-agent emotion)

$$\nvdash \mathit{Bel}_i\, \mathit{Admiration}_{j,i}(i{:}\alpha, \varphi) \rightarrow \mathit{Pride}_i\,(i{:}\alpha, \varphi) \quad \text{(a)}$$

$$\nvdash \mathit{Bel}_i\, \mathit{Reproach}_{j,i}(i{:}\alpha, \varphi) \rightarrow \mathit{Shame}_i\,(i{:}\alpha, \varphi) \quad \text{(b)}$$

Both prospect-based emotions and attribution emotions involve probabilities. We thus get interested in their temporal links with each other. We can prove that if an agent feels an attribution emotion about an action with a given result, and that before this action he envisaged that it could happen with this result and had a corresponding desire, then at this moment he felt a prospect-based emotion about the performance of this action with this result (namely about the success or failure of the action with respect to the prospected result). We have the same theorem if the agent feeling the emotion is different from the agent performing the action.

Theorem 8 (Link between prospect and attribution emotions)

$$\vdash \mathit{Pride}_i\,(i{:}\alpha, \varphi) \rightarrow \mathit{Bel}_i\, \mathit{Done}_{i:\alpha}\,((\neg \mathit{Bel}_i\, \neg \mathit{Happens}_{i:\alpha}\, \varphi \wedge \mathit{Des}_i\, \mathit{Happens}_{i:\alpha}\, \varphi) \rightarrow \mathit{Fear}_i\, \neg \mathit{Happens}_{i:\alpha}\, \varphi) \quad \text{(a)}$$

$$\vdash \mathit{Shame}_i\,(i{:}\alpha, \varphi) \rightarrow \mathit{Bel}_i\, \mathit{Done}_{i:\alpha}\,((\neg \mathit{Bel}_i\, \neg \mathit{Happens}_{i:\alpha}\, \varphi \wedge \mathit{Des}_i\, \neg \mathit{Happens}_{i:\alpha}\, \varphi) \rightarrow \mathit{Hope}_i\, \neg \mathit{Happens}_{i:\alpha}\, \varphi) \quad \text{(b)}$$

$$\vdash \mathit{Admiration}_{i,j}(j{:}\alpha, \varphi) \rightarrow \mathit{Bel}_i\, \mathit{Done}_{j:\alpha}\,((\neg \mathit{Bel}_i\, \neg \mathit{Happens}_{j:\alpha}\, \varphi \wedge \mathit{Des}_i\, \mathit{Happens}_{j:\alpha}\, \varphi) \rightarrow \mathit{Fear}_i\, \neg \mathit{Happens}_{j:\alpha}\, \varphi) \quad \text{(c)}$$

$$\vdash \mathit{Reproach}_{i,j}(j{:}\alpha, \varphi) \rightarrow \mathit{Bel}_i\, \mathit{Done}_{j:\alpha}\,((\neg \mathit{Bel}_i\, \neg \mathit{Happens}_{j:\alpha}\, \varphi \wedge \mathit{Des}_i\, \neg \mathit{Happens}_{j:\alpha}\, \varphi) \rightarrow \mathit{Hope}_i\, \neg \mathit{Happens}_{j:\alpha}\, \varphi) \quad \text{(d)}$$

We can notice that we have to impose that the agent had a corresponding desire in order to make him feel fear or hope. Moral values are not sufficient to trigger these emotions, since they can be inconsistent with desires. For example one can desire to kill someone he hates while his moral values tell him not to do so.

We can also prove a kind of converse of this theorem: if the agent fears (resp. hopes) that he does not perform the action α with result φ, and that this performance is ideal for him (resp. unideal), then after he performed α, if he believes that φ is true then he feels pride (resp. shame). Actually, the agent was afraid to fail (resp. he hoped to succeed). For example someone who passes an examination and has few chances to succeed would feel afraid of failing, and then if he succeeds he would feel pride because it was difficult.

Theorem 9 (Link between attribution and prospect emotions) *If α is an action that the agent i believes to influence the proposition φ (*cf. *Remark 2), then:*

$$\vdash \mathit{Fear}_i\, \neg \mathit{Happens}_{i:\alpha}\, \varphi \wedge \mathit{Bel}_i\, \mathit{Idl}_i\, \mathit{Happens}_{i:\alpha}\, \varphi \rightarrow \mathit{After}_{i:\alpha}\, (\mathit{Bel}_i\, \varphi \rightarrow \mathit{Pride}_i\, (i{:}\alpha, \varphi)) \quad \text{(a)}$$

$$\vdash \mathit{Hope}_i\, \neg \mathit{Happens}_{i:\alpha}\, \varphi \wedge \mathit{Bel}_i\, \mathit{Idl}_i\, \neg \mathit{Happens}_{i:\alpha}\, \varphi \rightarrow \mathit{After}_{i:\alpha}\, (\mathit{Bel}_i\, \varphi \rightarrow \mathit{Shame}_i\, (i{:}\alpha, \varphi)) \quad \text{(b)}$$

$$\vdash \mathit{Fear}_i\, \neg \mathit{Happens}_{j:\alpha}\, \varphi \wedge \mathit{Bel}_i\, \mathit{Idl}_j\, \neg \mathit{Happens}_{j:\alpha}\, \varphi \rightarrow \mathit{After}_{j:\alpha}\, (\mathit{Bel}_i\, \varphi \rightarrow \mathit{Admiration}_{i,j}(j{:}\alpha, \varphi)) \quad \text{(c)}$$

$$\vdash \mathit{Hope}_i\, \neg \mathit{Happens}_{j:\alpha}\, \varphi \wedge \mathit{Bel}_i\, \mathit{Idl}_j\, \neg \mathit{Happens}_{j:\alpha}\, \varphi \rightarrow \mathit{After}_{j:\alpha}\, (\mathit{Bel}_i\, \varphi \rightarrow \mathit{Reproach}_{i,j}(j{:}\alpha, \varphi)) \quad \text{(d)}$$

6.4 Inconsistencies between some emotions

We can prove several inconsistencies between pairs of emotions.

First, we can prove the inconsistency between opposite emotions about the same proposition (polar opposites), *viz.* between the positive and the negative emotion of the same group. This is in agreement with the psychological definitions.

Theorem 10 (Polar inconsistencies)

$$\vdash \neg(\mathit{Joy}_i\, \varphi \wedge \mathit{Fear}_i\, \varphi)$$
$$\vdash \neg(\mathit{Hope}_i\, \varphi \wedge \mathit{Fear}_i\, \varphi)$$
$$\vdash \neg(\mathit{Satisfaction}_i\, \varphi \wedge \mathit{FearConfirmed}_i\, \varphi)$$
$$\vdash \neg(\mathit{Relief}_i\, \varphi \wedge \mathit{Disappointment}_i\, \varphi)$$
$$\vdash \neg(\mathit{HappyFor}_{i,j}\varphi \wedge \mathit{SorryFor}_{i,j}\varphi)$$
$$\vdash \neg(\mathit{Resentment}_{i,j}\varphi \wedge \mathit{Gloating}_{i,j}\varphi)$$
$$\vdash \neg(\mathit{Pride}_i\, (i{:}\alpha, \varphi) \wedge \mathit{Shame}_i\, (i{:}\alpha, \varphi))$$
$$\vdash \neg(\mathit{Admiration}_{i,j}(j{:}\alpha, \varphi) \wedge \mathit{Reproach}_{i,j}(j{:}\alpha, \varphi))$$
$$\vdash \neg(\mathit{Gratification}_i\, (i{:}\alpha, \varphi) \wedge \mathit{Remorse}_i\, (i{:}\alpha, \varphi))$$
$$\vdash \neg(\mathit{Gratitude}_{i,j}(j{:}\alpha, \varphi) \wedge \mathit{Anger}_{i,j}(j{:}\alpha, \varphi))$$

This follows in particular from the rationality axioms (D) for our operators Bel_i , Des_i , Prob_i and Idl_i .

Please notice that we can still capture mixed emotions about a given event, because these mixed emotions actually concern different aspects of this event, that we represent with different formulas as if they were different consequences of the main event. For example one who loses a friend who suffered from a long and painful disease will feel sadness about the loss of his friend, and at the same time relief about the end of his friend's suffering. We thus consider that there are two appraised events: the loss of a friend, that is undesirable, and the end of his suffering, that is desirable. The initial event (the death of a friend) thus triggers a positive and a negative emotion.

Due to the properties of our probability operator, hope is not only inconsistent with fear about the same φ but also with fear about $\neg\varphi$. Actually, depending on which one is more probable between φ and $\neg\varphi$, the agent feels either hope or fear. Thus these two emotions cannot occur simultaneously.

Theorem 11 (Non simultaneity of hope and fear)

$$\vdash \neg(Hope_i\,\varphi \wedge Fear_i\,\neg\varphi)$$

This is because by definitions $Hope_i\,\varphi$ implies $Prob_i\,\varphi$ while $Fear_i\,\neg\varphi$ implies $Prob_i\,\neg\varphi$, which cannot simultaneously be the case due to the consistency of expectations (Property (5) of Section 4.2).

Moreover, an agent cannot feel simultaneously a good-will and an ill-will emotion towards the same agent about the same issue.

Theorem 12 (Inconsistency between good-will and ill-will emotions)

$$\vdash \neg(HappyFor_{i,j}\varphi \wedge Resentment_{i,j}\varphi) \quad \text{(a)}$$

$$\vdash \neg(SorryFor_{i,j}\varphi \wedge Gloating_{i,j}\varphi) \quad \text{(b)}$$

$$\vdash \neg(HappyFor_{i,j}\varphi \wedge Gloating_{i,j}\varphi) \quad \text{(c)}$$

$$\vdash \neg(SorryFor_{i,j}\varphi \wedge Resentment_{i,j}\varphi) \quad \text{(d)}$$

The proof follows from the rationality axioms for Bel_i and Des_i (see the appendix for details).

6.5 Other interesting properties

Our formalism allows us to prove that an agent is aware of his emotions.

Theorem 13 (Emotional awareness) *For every emotion $Emotion_i$ among the twenty emotions that we have defined:*

$$\vdash Emotion_i\varphi \leftrightarrow Bel_i\,Emotion_i\varphi \quad \text{(a)}$$

$$\vdash \neg Emotion_i\varphi \leftrightarrow Bel_i\,\neg Emotion_i\varphi \quad \text{(b)}$$

This follows in particular from the introspection axioms for our operators Bel_i , $Prob_i$ and $Expect_i$.

According to [48], only situations that are relevant to the individual's well-being can trigger an emotion. If we consider that an event is relevant to i's well-being when it involves one of i's desires or values, then this is in agreement with the following theorem. Indeed, if the agent has no desire or ideal at all then no event is relevant to him, and thus no situation can trigger an emotion. Besides, desires and moral values are part of what Lazarus calls "ego-involvement".

Theorem 14 (Emotions and ego-involvement) *An agent who has neither desires nor ideals cannot feel any emotion.*

The proof trivially follows from the definitions of emotions, that all necessarily entail either a desire (for the event-based ones) or an ideal (for the agent-based ones). Compound emotions entail both a desire and an ideal.

7 Discussion

Logical approaches of emotions are still quite rare. J.J. Meyer is one of the few researchers to have contributed to this field. In particular he has recently proposed an approach [58] where emotions are considered as kinds of events, and where definitions like those presented above are the necessary conditions of the triggering of these events (that Searle would call "mental events" [76, Chap. 3]). This model is a very interesting alternative to ours, independently from the details of the definitions respectively chosen in each approach.

[57] proposes a logical model of emotions based on KARO, his logic of action, belief and choice (*cf*. [82] or [59]). He uses this logic to write generation rules for four emotions: joy, sadness, anger and fear, depending on the agent's plans. First, the generation conditions of these emotions only depend on the satisfaction of the agent's plans, making this model task-oriented. Indeed, Meyer's aim, as he states himself[20], is not to be faithful to psychological definitions but to design artificial agents. On the contrary, in our work, we try to stay as close as possible to the original psychological definitions of the emotions that we formalize, through building on one of the most widely used approaches, namely Ortony, Clore, and Collins' typology. Second, this approach focuses on the individual aspects of emotions: as there is no operator to represent social standards, no social emotion like pride or shame can be represented. Finally, we thus provide an emotional formalism that is richer (with twenty emotions) and more faithful to psychology. However, our formalism is still limited to the triggering of emotions, whereas Meyer and colleagues already formalized the influence of emotions on the agents' plans [22].

We would now like to highlight the assets and limitations of our own model from several points of view, namely computer science, logic and psychology.

Our model undoubtedly suffers from some limitations. First, from the logical point of view our framework lacks some expressivity. In particular we preferred not to use the full collection of existing temporal operators like $Since$ or $Until$ in order to keep

[20] "Instead of trying to capture the informal psychological descriptions exactly (or as exact as possible), we primarily look here at a description that makes sense for artificial agents." [57, p.11]

our logic simple (see Footnote 19). As for our other choices, they are mainly due to the state of the art in BDI-like logics. First, in some places we had to approximate concepts. Most importantly, our account is incomplete as to the link between action and consequences, because propositional dynamic logic does not provide it (see Remark 2). Our logic therefore does not fully account for the notion of responsibility in agent-based emotions. Second, in some places we had to ignore concepts entirely because there is no logical operator in the literature that would allow to take them into account. Most importantly, the concept of goal does not appear in our logic, and in consequence its link with desirability is neglected. The reason here is that there exists no consensical logical analysis up to now.

Our emotions have no intensity degrees because it is not easy to design a semantics for graded operators and their evolution in time. We plan to further investigate this based on the logic of graded belief of [46]. However, despite all these limitations we believe that our formalism is expressive enough to give satisfying definitions of twenty emotions.

A point that is related to the previous one is that from the psychological point of view there are still several insufficiencies in our model. Mainly, our emotions are not quantitative: they have no intensity degree. This prevents us from fine-grained differentiations among emotions of the same type (for example: irritation, anger, rage). A second (and linked) shortcoming is that we do not manage the emotional dynamics: our emotions are persistent as long as their conditions stay true. Thereby some emotions (like Joy or $Satisfaction$) can persist *ad vitam eternam*, which is not intuitive at all. Indeed it has been established in psychology that after an emotion is triggered, its intensity decreases, and disappears when it is below a threshold. Finally, we cannot manage emotional blending of several emotions that are simultaneously triggered; [33] proposes an original solution to this issue. We leave these problems for further work.

Moreover we only provided a formalization of the OCC theory, that is far from being as popular in psychology as it is in computer science. It was necessary to choose one theory to begin with, but we believe that our BDI framework is expressive enough to formalize other psychological theories, all the more they often share the same appraisal variables. We already saw that we capture implicitly the control variable defined by [48]. On the contrary we do not capture the coping potential variable because it does not intervene in the triggering of OCC emotions; however we can represent it and we did so when formalizing coping strategies [4]. In this paper we only formalized the triggering of emotions, but this is the necessary starting point before formalizing their influence on any cognitive process. We neither formalize the subsequent life of emotions: their temporal decay (since we have no associated intensity degrees), and their interaction with mood or personality, but this is an interesting extension of this work.

From the logical point of view our model offers a clear semantics, which we think is quite rare in existing logical models of emotions. It also allows to highlight the power of BDI logics to reason about and disambiguate complex concepts, since we were able to prove some intuitive properties of emotions thanks to our logical definitions. It is only a logical formalism can give such unequivocal results about phenomena that are not always clearly analyzed in the psychological literature. Finally our model somehow validates BDI logics, that were designed to formalize mental attitudes, since it demonstrates that they are expressive enough to characterize as complex mental attitudes as

emotions (we recall that philosophers like Searle consider emotions as complex mental attitudes, see Section 2.1).

From the psychologist's point of view, it could be interesting to validate theories thanks to the reasoning services offered by logic. Another way for them to take advantage from such a model is to conduct experiments with emotional agents endowed with it, instead of humans that are not always able to clearly analyze their own emotions. We have implemented such an emotional agent and have experimented it with human users analyzing the believability of its emotions. We then used the results of this experiments to derive conclusions about the underlying OCC theory. We have not conducted this work in collaboration with psychologists, but plan to do so in the near future.

Finally, from the computer science point of view this cross-disciplinary work brings an interesting contribution, since it fills the gap between psychology and the agent community. We designed a domain-independent model, based on a standard formalism, BDI logics, that are already used in a great number of agent architectures. Our model is thus ready to be implemented in any BDI agent, whatever its application may be. It will save designers the long and costly (though necessary) process of formalization of a psychological theory. Moreover it offers them a rich set of emotions that will make their agents very expressive. To illustrate all these assets we have ourselves implemented such an agent endowed with our model, only making some concessions, for example in order to add intensity degrees to emotions.

8 Conclusion

In this paper, we have formalized twenty emotions from OCC theory (all but the object-based branch), thus providing a very rich set of emotions. Moreover we have shown the soundness of our framework by illustrating each definition by an example from Ortony *et al.*'s book. We managed to formalize nearly the whole OCC theory with our logic, supporting the author's assumption that their theory is computationally tractable. On the contrary some appraisal variables from other theories, like Lazarus' ego-involvement, seem to be much more ambiguous and difficult to formalize.

We have privileged richness, genericity, and fidelity to the definitions over tractability. An optimization would have needed important concessions. For example [64] propose a numerical model of emotions in combat games, efficient in big real-time multi-agent systems, but which is domain-dependent.

In other works we have explored some extensions of this model. First we have provided an account of the influence of emotions on the agent's behavior by formalizing in the same BDI framework some coping strategies. According to psychologists [49], appraisal and coping are indivisible. However, the formalization of each process was a full-fledged work and we thus investigated them in separate papers. Our formalization of coping strategies [4] allows to explain how an agent modifies his beliefs or intentions depending on his current emotion.

Second, we have implemented our logical model of both appraisal and coping in a BDI agent [2]. This agent answers emotionally to stimuli sent by the user through the interface. This work is still in progress to implement other kinds of influence that emotions have on the agent: interaction with personality, modification of the reasoning

strategies (in the sense of [31]), impact on the agent's centers of attention (in the sense of the activation notion of [6])...

Our future research will be oriented towards several different aims. First we would like to use this logical framework to formalize various existing psychological theories of emotions. Once expressed in the same language, we would be able to compare these theories. Second we would like to conduct new experiments with our BDI agents, but this time in cooperation with psychologists who could help us interpreting the results.

From a logical perspective we will further investigate the links between mental attitudes, in particular how desirability can be computed from goals. Moreover, our work currently excludes object-based emotions: in future work a modal predicate logic could allow to characterize the properties of objects and thus define the emotions triggered by their appraisal. Finally, we might unify the formalization of events and actions by moving from dynamic-logic actions to theories of agency such as STIT theory or the logic of 'brining-it-about'.

To conclude, our cross-disciplinary approach combines the advantages of logic and computational models with the expertise of psychology of emotions. Even if the resulting computational model of emotions still suffers from some limitations, we hope that it already brings some interesting contributions for computer science and logic as well as for psychology itself.

Acknowledgements

We would like to thank the 3 reviewers for their thorough comments, which helped to improve the paper in several places.

References

[1] C. Adam, B. Gaudou, D. Longin, and E. Lorini. Logical modeling of emotions for ambient intelligence. *Journal of Ambient Intelligence and Smart environments*, in submission, 2009. Special issue "Contribution of AI to Ambient Intelligence".

[2] Carole Adam. *Emotions: from psychological theories to logical formalization and implementation in a BDI agent*. PhD thesis, INP Toulouse, France, July 2007. available in English.

[3] Carole Adam, Benoit Gaudou, Andreas Herzig, and Dominique Longin. OCC's emotions: a formalization in a BDI logic. In Jérôme Euzenat, editor, *Proc. of the Twelfth Int. Conf. on Artificial Intelligence: Methodology, Systems, and Applications (AIMSA'06), Varna, Bulgaria, september 13–15*, volume 4183 of *LNAI*, pages 24–32. Springer-Verlag, 2006.

[4] Carole Adam and Dominique Longin. Endowing emotional agents with coping strategies: from emotions to emotional behaviour. In C. Pelachaud *et al.*, editor, *Intelligent Virtual Agents (IVA'07)*, volume 4722 of *LNCS*, pages 348–349. Springer-Verlag, 2007.

[5] G. Aist, B. Kort, R. Reilly, J. Mostow, and R.W. Picard. Experimentally augmenting an intelligent tutoring system with human-supplied capabilities: Adding human-provided emotional scaffolding to an automated reading tutor that listens. In *Proceedings of the Fourth IEEE International Conference on Multimodal Interfaces (ICMI 2002)*, pages 483–490. IEEE Computer Society, 2002.

[6] J.R. Anderson and C. Lebiere. *The Atomic Components of Thought*. Lawrence Erlbaum Associates, Mahwah, NJ, 1998.

[7] L. Åqvist. Deontic Logic. In D.M. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic*, volume 8, pages 147–264. Kluwer Academic Publishers, 2nd edition, 2002.

[8] M.B. Arnold. *Emotion and personality*. Columbia University Press, New York, 1960.

[9] C. Bartneck. *eMuu—An Embodied Emotional Character for the Ambient Intelligent Home*. PhD thesis, Eindhoven University of Technology, 2002.

[10] J. Bates. The role of emotion in believable agents. *Communications of the ACM*, 37(7):122–125, 1994.

[11] Christian Becker, Stefan Kopp, and Ipke Wachsmuth. Simulating the emotion dynamics of a multimodal conversational agent. In *ADS'04*. Springer LNCS, 2004.

[12] Gordon H. Bower. How might emotions affect learning. In Sven-Ake Cristianson, editor, *The handbook of emotion and memory: research and theory*, pages 3–32. Lawrence Erlbaum Associates, 1992.

[13] Michael E. Bratman. *Intention, Plans, and Practical Reason*. Harvard University Press, Cambridge, MA, USA, 1987.

[14] S. Brave, C. Nass, and K. Hutchinson. Computers that care: Investigating the effects of orientation of emotion exhibited by an embodied computer agent. *International Journal of Human-Computer Studies*, 62:161–178, 2005.

[15] J. P. Burgess. Probability logic. *J. of Symbolic Logic*, 34:264–274, 1969.

[16] John P. Burgess. Basic tense logic. In Dov Gabbay and Franz Guenthner, editors, *Handbook of Philosophical Logic*, volume 7, pages 1–42. Kluwer Academic Publishers, 2nd edition, 2002.

[17] Cristiano Castelfranchi and Fabio Paglieri. The role of belief in goal dynamics: Prolegomena to a constructive theory of intentions. *Synthese*, to appear, 2007.

[18] B. F. Chellas. *Modal Logic: an Introduction*. Cambridge University Press, Cambridge, 1980.

[19] Philip R. Cohen and Hector J. Levesque. Intention is choice with commitment. *Artificial Intelligence Journal*, 42(2–3):213–261, 1990.

[20] Antonio R. Damasio. *Descartes' Error: Emotion, Reason, and the Human Brain*. Putnam Pub Group, 1994.

[21] Charles R. Darwin. *The expression of emotions in man and animals*. Murray, London, 1872.

[22] Mehdi Dastani and John-Jules Meyer. Programming agents with emotions. In *Proceedings 17th European Conf. on Artificial Intelligence (ECAI 2006), Trento, Italy, Aug. 28th–Sep. 1st*. IOS Press, 2006.

[23] F. de Rosis, C. Pelachaud, I. Poggi, V. Carofiglio, and B. D. Carolis. From Greta's mind to her face: modelling the dynamics of affective states in a conversational embodied agent. *International Journal of Human-Computer Studies*, 59(1-2):81–118, 2003.

[24] P. Ekman, W.V. Friesen, and J.C. Hager. *Facial Action Coding System Invetigator's Guide*. A Human Face, 2002.

[25] Paul Ekman. An argument for basic emotions. *Cognition and Emotion*, 6:169–200, 1992.

[26] Clark Elliott. *The Affective Reasoner : A process model of emotions in a multi-agent system*. PhD thesis, Northwestern University, Illinois, 1992.

[27] Clark Elliott, Jeff Rickel, and James Lester. Lifelike pedagogical agents and affective computing: An exploratory synthesis. *Lecture Notes in Computer Science*, 1600:195–211, 1999.

[28] J. Elster. Emotions and economic theory. *Journal of Economic Literature*, 36(1):47–74, 1998.

[29] Ronald Fagin and Joseph Y. Halpern. Reasoning about knowledge and probability. *Journal of the ACM*, 41(2), 1994.

[30] Ronald Fagin, Joseph Y. Halpern, Moshe Y. Vardi, and Yoram Moses. *Reasoning about knowledge*. MIT Press, Cambridge, 1995.

[31] J.P. Forgas. Mood and judgment: The affect infusion model (aim). *Psychological Bulletin*, 117:39–66, 1995.

[32] N. H. Frijda. *The Emotions*. Cambridge University Press, 1986.

[33] Carlos Gershenson. Modelling emotions with multidimensional logic. In *NAFIPS'99*. IEEE, 1999.

[34] R. Gordon. *The structure of emotions*. Cambridge University Press, New York, 1987.

[35] J. Gratch and S. Marsella. A domain independent framework for modeling emotion. *Journal of Cognitive Systems Research*, 5(4):269–306, 2004.

[36] A. Grizard and C.L. Lisetti. Generation of facial emotional expressions based on psychological theory. In *Workshop on Emotion and Computing, KI'2006*, Bremen, Germany, June, 14-19 2006.

[37] J. Halpern and D. McAllester. Likelihood, probability, and knowledge. *Computational Intelligence*, 5, 1989.

[38] J. Halpern and M. Rabin. A logic to reason about likelihood. *Artificial Intelligence J.*, 32(3):379–405, 1987.

[39] Andreas Herzig. Modal probability, belief, and actions. *Fundamenta Informaticæ*, 57(2-4):323–344, 2003.

[40] Andreas Herzig and Dominique Longin. Sensing and revision in a modal logic of belief and action. In F. van Harmelen, editor, *Proc. of 15th European Conf. on Artificial Intelligence (ECAI 2002), Lyon, France, July 23–26*, pages 307–311. IOS Press, 2002.

[41] Andreas Herzig and Dominique Longin. C&L intention revisited. In Didier Dubois, Chris Welty, and Mary-Anne Williams, editors, *Proc. 9th Int. Conf. on Principles of Knowledge Representation and Reasoning (KR 2004), Whistler, Canada, June 2–5*, pages 527–535. AAAI Press, 2004.

[42] J. Hintikka. *Knowledge and Belief: An Introduction to the Logic of the Two Notions*. Cornell University Press, Ithaca, 1962.

[43] Patricia A. Jaques, Rosa M. Vicari, Sylvie Pesty, and Jean-Francois Bonneville. Applying affective tactics for a better learning. In *In Proceedings of the 16th European Conference on Artificial Intelligence (ECAI 2004)*. IOS Press, 2004.

[44] J. Klein, Y. Moon, and R. Picard. This computer responds to user frustration. In *Proceedings of the Conference on Human Factors in Computing Systems*, pages 242–243, Pittsburgh, USA, may 1999. ACM Press.

[45] Jérôme Lang, Leendert W. N. van Der Torre, and Emil Weydert. Utilitarian desires. *Journal of Autonomous Agents and Multi-Agent Systems*, 5:329–363, 2002.

[46] Noel Laverny and Jérôme Lang. From knowledge-based programs to graded belief-based programs, Part II: off-line reasoning . In *Proceedings of the 9th International Joint Conference on Artificiel Intelligence (IJCAI'05) , Edinburgh, Scotland, 31/07/05-05/08/05*, pages 497–502. Gallus, juillet 2005.

[47] Richard Lazarus. The Cognition–Emotion Debate: a Bit of History. In Tim Dalgleish and Mick Power, editors, *Handbook of Cognition and Emotion*, pages 3–20. John Wiley & Sons, New York, 1999.

[48] Richard S. Lazarus. *Emotion and Adaptation*. Oxford University Press, 1991.

[49] Richard S. Lazarus and Susan Folkman. *Stress, Appraisal, and Coping*. Springer Publishing Company, 1984.

[50] W. Lenzen. *Recent work in epistemic logic*. North Holland Publishing Company, Amsterdam, 1978.

[51] W. Lenzen. On the semantics and pragmatics of epistemic attitudes. In A. Laux and H. Wansing, editors, *Knowledge and belief in philosophy and AI*, pages 181–197. Akademie Verlag, Berlin, 1995.

[52] G. Loewenstein. Emotions in economic theory and economic behavior. *American Economic Review*, 90(2):426–432, 2000.

[53] E. Lorini and C. Castelfranchi. The unexpected aspects of surprise. *International Journal of Pattern Recognition and Artificial Intelligence*, 20(6):817–833, 2006.

[54] E. Lorini and C. Castelfranchi. The cognitive structure of surprise: looking for basic principles. *Topoi: an International Review of Philosophy*, forthcoming, 2007.

[55] Emiliano Lorini and Andreas Herzig. A logic of intention and attempt. *Synthese*, 163(1):45–77, 2008.

[56] J. J. Ch. Meyer, W. van der Hoek, and B. van Linder. A logical approach to the dynamics of commitments. *Artificial Intelligence*, 113(1–2):1–40, 1999.

[57] John Jules Meyer. Reasoning about emotional agents. In R. López de Mántaras and L. Saitta, editors, *16th European Conf. on Artif. Intell. (ECAI)*, pages 129–133, 2004.

[58] John-Jules Meyer. Reasoning about emotional agents. *International Journal of Intelligent Systems*, 21(6):601–619, 2006.

[59] John-Jules Meyer, Frank de Boer, Rogier van Eijk, Koen Hindriks, and Wiebe van der Hoek. On programming karo agents. *Logic Journal of the IGPL*, 9(22):261–272, 2001.

[60] W. U. Meyer, R. Reisenzein, and A. Schützwohl. Towards a process analysis of emotions: The case of surprise. *Motivation and Emotion*, 21:251–274, 1997.

[61] Magali Ochs, R. Niewiadomski, Catherine Pelachaud, and David Sadek. Intelligent expressions of emotions. In *1st International Conference on Affective Computing and Intelligent Interaction ACII*, China, October 2005.

[62] Andrew Ortony, G.L. Clore, and A. Collins. *The cognitive structure of emotions*. Cambridge University Press, Cambridge, MA, 1988.

[63] T. Partala and V. Surakka. The effects of affective interventions in human-computer interaction. *Interacting with computers*, 16:295–309, 2004.

[64] H. van Dyke Parunak, Robert Bisson, Sven Brueckner, Robert Matthews, and John Sauter. A model of emotions for situated agents. In Peter Stone and Gerhard Weiss, editors, *AAMAS'06*, pages 993–995. ACM Press, 2006.

[65] R.W. Picard and K.K. Liu. Relative Subjective Count and Assessment of Interruptive Technologies Applied to Mobile Monitoring of Stress. *International Journal of Human-Computer Studies*, 65:396–375, 2007.

[66] H. Prendinger and M. Ishizuka. The empathic companion: A character-based interface that addresses users' affective states. *International Journal of Applied Artificial Intelligence*, 19:297–285, 2005.

[67] Anand S. Rao and Michael P. Georgeff. Modeling rational agents within a BDI-architecture. In J. A. Allen, R. Fikes, and E. Sandewall, editors, *Proceedings Second Int. Conf. on Principles of Knowledge Representation and Reasoning (KR'91)*, pages 473–484. Morgan Kaufmann Publishers, 1991.

[68] Anand S. Rao and Michael P. Georgeff. An abstract architecture for rational agents. In Bernhard Nebel, Charles Rich, and William Swartout, editors, *Proceedings Third Int. Conf. on Principles of Knowledge Representation and Reasoning (KR'92)*, pages 439–449. Morgan Kaufmann Publishers, 1992.

[69] Neal Reilly. *Believable social and emotional agents*. PhD thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA, 1996.

[70] Ray Reiter. The frame problem in the Situation Calculus: A simple solution (sometimes) and a completeness result for goal regression. In Vladimir Lifschitz, editor, *Artificial Intelligence and Mathematical Theory of Computation: Papers in Honor of John McCarthy*, pages 359–380. Academic Press, San Diego, CA, 1991.

[71] James A. Russell. How shall an emotion be called? In R. Plutchik and H.R. Conte, editors, *Circumplex models of personality and emotions*, pages 205–220. American Psychological Association, Washington, DC, 1997.

[72] M.D. Sadek. A study in the logic of intention. In Bernhard Nebel, Charles Rich, and William Swartout, editors, *Proceedings Third Int. Conf. on Principles of Knowledge Representation and Reasoning (KR'92)*, pages 462–473. Morgan Kaufmann Publishers, 1992.

[73] K. R. Scherer. *Appraisal Processes in Emotion : Theory, Methods, Research*, chapter Appraisal Considered as a Process of Multilevel Sequential Checking, pages 92–120. Oxford University Press, New York, 2001.

[74] K.R. Scherer. Toward a dynamic theory of emotion: the component process model of affective states. *Geneva studies in Emotion and Communication*, 1(1):1–98, 1987.

[75] J. R. Searle. *Speech acts: An essay in the philosophy of language*. Cambridge University Press, New York, 1969.

[76] John R. Searle. *Intentionality: An essay in the philosophy of mind*. Cambridge University Press, 1983.

[77] K. Segerberg. Qualitative probability in a modal setting. In J. Fenstad, editor, *Proceedings of the 2nd Scandinavian Logic Symp.*, Amsterdam, 1971. North Holland Publ. Company.

[78] P. Shaver, J. Schwartz, D. Kirson, and C. O'Connor. Emotion knowledge. *Journal of personality and Social Psychology*, 52, 1987.

[79] R. Solomon and C. Calhoun, editors. *What is an Emotion? Classic Readings in Philosophical Psychology*. Oxford University Press, Oxford, 1984.

[80] Alexander Staller and Paolo Petta. Introducing emotions into the computational study of social norms: a first evaluation. *Journal of Artificial Societies and Social Simulation*, 4(1), 2001.

[81] B. R. Steunebrink, M. Dastani, and J. J. Ch. Meyer. A logic of emotions for intelligent agents. In *Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence (AAAI'07)*, pages 142–147. AAAI Press, 2007.

[82] W. van der Hoek, B. van Linder, and J-J. Ch. Meyer. An integrated modal approach to rational agents. In *Proceedings of 2nd AISB Workshop on Practical Reasoning and Rationality*, pages 123–159, Manchester, United Kingdom, April 7-9 1997.

[83] P. Walley and T. L. Fine. Varieties of modal (classificatory) and comparative probability. *Synthese*, 41, 1979.

In order not to overload the paper, we gather in this appendix the proofs of the theorems given in the main part. This appendix is intended to help the reviewer to understand the theorems. It may be dropped in the final version of the paper.

In the proofs, $\mathcal{PL}$ refers to the Propositional Logic, and $\mathcal{ML}$ refers to the principles of normal modal logic.

Theorem 1 (Temporal link from prospect to confirmation)

$$\vdash Bel_i\, PHope_i\, \varphi \wedge (Bel_i\, \varphi \vee Bel_i\, \neg\varphi) \rightarrow Satisfaction_i\, \varphi \vee Disappointment_i\, \neg\varphi \quad \text{(a)}$$

$$\vdash Bel_i\, PFear_i\, \varphi \wedge (Bel_i\, \varphi \vee Bel_i\, \neg\varphi) \rightarrow Relief_i\, \varphi \vee FearConfirmed_i\, \neg\varphi \quad \text{(b)}$$

To prove the Theorem 1 we need the following lemma.

Lemma 1 $\vdash Bel_i\, PDes_i\, \varphi \rightarrow Des_i\, \varphi$.

Proof 1 (of Lemma 1)

1. $\vdash Des_i\, \varphi \rightarrow GDes_i\, \varphi$ (**from** (Pers-Des_i))
2. $\vdash PDes_i\, \varphi \rightarrow PGDes_i\, \varphi$ (**from** 1. by $\mathcal{ML}$)
3. $\vdash PGDes_i\, \varphi \rightarrow Des_i\, \varphi$ (**from** (CONV-HF) by $\mathcal{PL}$)
4. $\vdash PDes_i\, \varphi \rightarrow Des_i\, \varphi$ (**from** 2. and 3. by $\mathcal{PL}$)
5. $\vdash Bel_i\, PDes_i\, \varphi \rightarrow Bel_i\, Des_i\, \varphi$ (**from** 4. by (RM-$\Box$) for Bel_i)
6. $\vdash Bel_i\, Des_i\, \varphi \rightarrow Des_i\, \varphi$ (**from** (5-MIX2) and (D-Bel_i))
7. $\vdash Bel_i\, PDes_i\, \varphi \rightarrow Des_i\, \varphi$ (**from** 5. and 6. by $\mathcal{PL}$)

□

Proof 2 (of Theorem 1) *Case of (a). Actually it suffices to prove that (i)* $Bel_i\, PHope_i\, \varphi \wedge Bel_i\, \varphi \rightarrow Satisfaction_i\, \varphi$ *and (ii)* $Bel_i\, PHope_i\, \varphi \wedge Bel_i\, \neg\varphi \rightarrow Disappointment_i\, \neg\varphi$ *are theorems. Case of (i).*

1. $\vdash Bel_i\, PHope_i\, \varphi \rightarrow Bel_i\, P(Expect_i\, \varphi \wedge Des_i\, \varphi)$ (**from** definition 1)
2. $\vdash Bel_i\, PHope_i\, \varphi \rightarrow Bel_i\, PExpect_i\, \varphi \wedge Bel_i\, PDes_i\, \varphi$ (**by** $\mathcal{ML}$)
3. $\vdash Bel_i\, PHope_i\, \varphi \rightarrow Bel_i\, PExpect_i\, \varphi \wedge Des_i\, \varphi$ (**by** Lemma 1)
4. $\vdash Bel_i\, PHope_i\, \varphi \wedge Bel_i\, \varphi \rightarrow Bel_i\, PExpect_i\, \varphi \wedge Des_i\, \varphi \wedge Bel_i\, \varphi$ (**by** $\mathcal{PL}$)
5. $\vdash Bel_i\, PHope_i\, \varphi \wedge Bel_i\, \varphi \rightarrow Satisfaction_i\, \varphi$ (**by** def. of $Satisfaction$)

We demonstrate (ii) in the same way. Case of (b): the proof is similar. □

Theorem 3 (Link between confirmation and well-being emotions)

$$\vdash Satisfaction_i\,\varphi \rightarrow Joy_i\,\varphi \quad \text{(a)}$$
$$\vdash FearConfirmed_i\,\varphi \rightarrow Distress_i\,\varphi \quad \text{(b)}$$
$$\vdash Relief_i\,\varphi \rightarrow Joy_i\,\varphi \quad \text{(c)}$$
$$\vdash Disappointment_i\,\varphi \rightarrow Distress_i\,\varphi \quad \text{(d)}$$

Proof 3 (of Theorem 3) *Case of (a).*

1. $\vdash Satisfaction_i\,\varphi \rightarrow Bel_i\,\varphi \wedge Des_i\,\varphi$ (**from** def. of $Satisfaction$)
2. $\vdash Satisfaction_i\,\varphi \rightarrow Joy_i\,\varphi$ (**by** definition of Joy)

The proof is similar for cases (b) to (d). □

Theorem 4 (From fortune-of-other emotion to image of other)

$$\vdash HappyFor'_{i,j}\varphi \rightarrow Prob_i\,F\,Joy_j\,\varphi \quad \text{(a)}$$
$$\vdash SorryFor'_{i,j}\varphi \rightarrow Prob_i\,F\,Distress_j\,\varphi \quad \text{(b)}$$
$$\vdash Resentment'_{i,j}\varphi \rightarrow Prob_i\,F\,Joy_j\,\varphi \quad \text{(c)}$$
$$\vdash Gloating'_{i,j}\varphi \rightarrow Prob_i\,F\,Distress_j\,\varphi \quad \text{(d)}$$

Proof 4 (of Theorem 4) *Case of (a).*

1. $\vdash HappyFor'_{i,j}\varphi \rightarrow Prob_i\,F\,Bel_j\,\varphi \wedge Bel_i\,Des_j\,\varphi$ (**from** definition of $HappyFor'$)
2. $\vdash HappyFor'_{i,j}\varphi \rightarrow Prob_i\,(F\,Bel_j\,\varphi \wedge G\,Des_j\,\varphi)$ (**by** (Pers-Des_i) and (C-MIX))
3. $\vdash HappyFor'_{i,j}\varphi \rightarrow Prob_i\,F(Bel_j\,\varphi \wedge Des_j\,\varphi)$ (**by** property (1) for G)
4. $\vdash HappyFor'_{i,j}\varphi \rightarrow Prob_i\,F\,Joy_j\,\varphi$ (**by** definition of Joy)

The proof is similar for cases (b) to (d). □

Theorem 5 (Consequences of fortunes-of-others emotions)

$$\vdash HappyFor'_{i,j}\varphi \wedge \neg Bel_i\,F\,Bel_j\,\varphi \rightarrow Hope_i\,F\,Bel_j\,\varphi \quad \text{(a)}$$
$$\vdash SorryFor'_{i,j}\varphi \wedge \neg Bel_i\,F\,Bel_j\,\varphi \rightarrow Fear_i\,F\,Bel_j\,\varphi \quad \text{(b)}$$
$$\vdash Resentment'_{i,j}\varphi \wedge \neg Bel_i\,F\,Bel_j\,\varphi \rightarrow Fear_i\,F\,Bel_j\,\varphi \quad \text{(c)}$$
$$\vdash Gloating'_{i,j}\varphi \wedge \neg Bel_i\,F\,Bel_j\,\varphi \rightarrow Hope_i\,F\,Bel_j\,\varphi \quad \text{(d)}$$

Proof 5 (of Theorem 5) *Case of (a).*

1. $\vdash HappyFor'_{i,j}\varphi \rightarrow Prob_i\, F Bel_j\, \varphi \wedge Des_i\, Bel_j\, \varphi$
(**from** definition of $HappyFor'$)

2. $\vdash HappyFor'_{i,j}\varphi \rightarrow Prob_i\, F Bel_j\, \varphi \wedge Des_i\, F Bel_j\, \varphi$
(**by** contraposition of (T-G), and (RM-$\Box$) for Des_i)

3. $\vdash HappyFor'_{i,j}\varphi \wedge \neg Bel_i\, F Bel_j\, \varphi \rightarrow$
$Prob_i\, F Bel_j\, \varphi \wedge \neg Bel_i\, F Bel_j\, \varphi \wedge Des_i\, F Bel_j\, \varphi$ (**by** $\mathcal{PL}$)

4. $\vdash HappyFor'_{i,j}\varphi \wedge \neg Bel_i\, F Bel_j\, \varphi \rightarrow$
$Expect_i\, F Bel_j\, \varphi \wedge Des_i\, F Bel_j\, \varphi$ (**by** definition 1)

5. $\vdash HappyFor'_{i,j}\varphi \wedge \neg Bel_i\, F Bel_j\, \varphi \rightarrow Hope_i\, F Bel_j\, \varphi$
(**by** definition of $Hope$)

The proof is similar for cases (b) to (d). □

Theorem 6 (Other-agent emotions towards oneself)

$$\vdash Admiration_{i,i}(i{:}\alpha, \varphi) \leftrightarrow Pride_i\,(i{:}\alpha, \varphi) \quad \text{(a)}$$
$$\vdash Reproach_{i,i}(i{:}\alpha, \varphi) \leftrightarrow Shame_i\,(i{:}\alpha, \varphi) \quad \text{(b)}$$

Proof 6 (of Theorem 6) *Case of (a). The proof comes immediately from the definitions of these two emotions.*

1. $\vdash Admiration_{i,i}(i{:}\alpha, \varphi) \leftrightarrow Bel_i\, Done_{i:\alpha}\,(\neg Prob_i\, Happens_{i:\alpha}\, \top \wedge$
$Bel_i\, Idl_i\, Happens_{i:\alpha}\, \top)$ (**by** definition of $Admiration$)

2. $\vdash Admiration_{i,i}(i{:}\alpha, \varphi) \leftrightarrow Pride_i\,(i{:}\alpha, \varphi)$ (**by** definition of $Pride$)

The proof is similar for (b). □

Theorem 7 (Other-agent emotion does not force self-agent emotion)

$$\not\vdash Bel_i\, Admiration_{j,i}(i{:}\alpha, \varphi) \rightarrow Pride_i\,(i{:}\alpha, \varphi) \quad \text{(a)}$$
$$\not\vdash Bel_i\, Reproach_{j,i}(i{:}\alpha, \varphi) \rightarrow Shame_i\,(i{:}\alpha, \varphi) \quad \text{(b)}$$

Sketch of proof 1 (of Theorem 7) *It suffices to find a counter-example,* viz. *a model where the implication is not valid,* viz. *a model containing at least one world where the implication is false.*

Case of (b). By definition, $Bel_j\, Reproach_{i,j}(j{:}\alpha, \varphi)$ *does not imply* $Des_j\, \neg Happens_{j:\alpha}\, \varphi$. *In a world where the first formula is true and the second one is false, the implication is false. For example, a teacher in a school can reproach to a student to wear unauthorised clothes, and tell this to him, without making this student ashamed of wearing them.*

Theorem 8 (Link between prospect and attribution emotions)

$$\vdash Pride_i\,(i{:}\alpha, \varphi) \rightarrow Bel_i\, Done_{i:\alpha}\,((\neg Bel_i\, \neg Happens_{i:\alpha}\,\varphi \wedge Des_i\, Happens_{i:\alpha}\,\varphi) \rightarrow Fear_i\, \neg Happens_{i:\alpha}\,\varphi) \quad \text{(a)}$$

$$\vdash Shame_i\,(i{:}\alpha, \varphi) \rightarrow Bel_i\, Done_{i:\alpha}\,((\neg Bel_i\, \neg Happens_{i:\alpha}\,\varphi \wedge Des_i\, \neg Happens_{i:\alpha}\,\varphi) \rightarrow Hope_i\, \neg Happens_{i:\alpha}\,\varphi) \quad \text{(b)}$$

$$\vdash Admiration_{i,j}(j{:}\alpha, \varphi) \rightarrow Bel_i\, Done_{j:\alpha}\,((\neg Bel_i\, \neg Happens_{j:\alpha}\,\varphi \wedge Des_i\, Happens_{j:\alpha}\,\varphi) \rightarrow Fear_i\, \neg Happens_{j:\alpha}\,\varphi) \quad \text{(c)}$$

$$\vdash Reproach_{i,j}(j{:}\alpha, \varphi) \rightarrow Bel_i\, Done_{j:\alpha}\,((\neg Bel_i\, \neg Happens_{j:\alpha}\,\varphi \wedge Des_i\, \neg Happens_{j:\alpha}\,\varphi) \rightarrow Hope_i\, \neg Happens_{j:\alpha}\,\varphi) \quad \text{(d)}$$

Proof 7 (of Theorem 8) *Case of (a).*

1. $\vdash Pride_i\,(i{:}\alpha, \varphi) \rightarrow Bel_i\, Done_{i:\alpha}\,(Prob_i\, After_{i:\alpha}\,\neg\varphi)$
   (**by** definition of $Pride$)
2. $\vdash Pride_i\,(i{:}\alpha, \varphi) \rightarrow Bel_i\, Done_{i:\alpha}\,(Prob_i\, \neg Happens_{i:\alpha}\,\varphi)$
   (**by** definition of $Happens$)
3. $\vdash Pride_i\,(i{:}\alpha, \varphi) \rightarrow Bel_i\, Done_{i:\alpha}\,(\neg Bel_i\, \neg Happens_{i:\alpha}\,\varphi \rightarrow Expect_i\, \neg Happens_{i:\alpha}\,\varphi)$ (**by** $\mathcal{PL}$ and definition 1)
4. $\vdash Pride_i\,(i{:}\alpha, \varphi) \rightarrow Bel_i\, Done_{i:\alpha}\,(\neg Bel_i\, \neg Happens_{i:\alpha}\,\varphi \wedge Des_i\, Happens_{i:\alpha}\,\varphi \rightarrow Fear_i\, \neg Happens_{i:\alpha}\,\varphi)$
   (**by** $\mathcal{PL}$ and definition of $Fear$)

The proof is similar for (b), (c) and (d). □

Theorem 9 (Link between attribution and prospect emotions) *If α is an action that the agent i believes to influence the proposition φ (*cf. *remark 2), then:*

$$\vdash Fear_i\, \neg Happens_{i:\alpha}\,\varphi \wedge Bel_i\, Idl_i\, Happens_{i:\alpha}\,\varphi \rightarrow After_{i:\alpha}\,(Bel_i\,\varphi \rightarrow Pride_i\,(i{:}\alpha, \varphi)) \quad \text{(a)}$$

$$\vdash Hope_i\, \neg Happens_{i:\alpha}\,\varphi \wedge Bel_i\, Idl_i\, \neg Happens_{i:\alpha}\,\varphi \rightarrow After_{i:\alpha}\,(Bel_i\,\varphi \rightarrow Shame_i\,(i{:}\alpha, \varphi)) \quad \text{(b)}$$

$$\vdash Fear_i\, \neg Happens_{j:\alpha}\,\varphi \wedge Bel_i\, Idl_j\, \neg Happens_{j:\alpha}\,\varphi \rightarrow After_{j:\alpha}\,(Bel_i\,\varphi \rightarrow Admiration_{i,j}(j{:}\alpha, \varphi)) \quad \text{(c)}$$

$$\vdash Hope_i\, \neg Happens_{j:\alpha}\,\varphi \wedge Bel_i\, Idl_j\, \neg Happens_{j:\alpha}\,\varphi \rightarrow After_{j:\alpha}\,(Bel_i\,\varphi \rightarrow Reproach_{i,j}(j{:}\alpha, \varphi)) \quad \text{(d)}$$

To prove Theorem 9 we need the following lemma.

Lemma 2 $Done_\alpha\, \neg Bel_i\, After_\alpha\, \bot \wedge Done_\alpha\, Bel_i\,\varphi \rightarrow Bel_i\, Done_\alpha\,\varphi$

To prove Lemma 2 we need the following lemma.

Lemma 3 *if* $\varphi \rightarrow \mathit{After}_\alpha\, \psi$ *then* $\mathit{Done}_\alpha\, \varphi \rightarrow \psi$

Proof 8 (of Lemma 3)

1. $\varphi \rightarrow \mathit{After}_\alpha\, \psi$ (**by** hypothesis)
2. $\mathit{Done}_\alpha\, \mathit{After}_\alpha\, \varphi \rightarrow \varphi$ (**from** contraposition of (CONV-BH))
3. $\mathit{Done}_\alpha\, \varphi \rightarrow \mathit{Done}_\alpha\, \mathit{After}_\alpha\, \psi$ (**from** 1. by (RM-$\Diamond$) for Done_α)
4. $\mathit{Done}_\alpha\, \varphi \rightarrow \psi$ (**from** 2. and 3.) $\Box$

Proof 9 (of Lemma 2)

1. $\mathit{Bel}_i\, \mathit{After}_\alpha\, \varphi \wedge \neg \mathit{Bel}_i\, \mathit{After}_\alpha\, \bot \rightarrow \mathit{After}_\alpha\, \mathit{Bel}_i\, \varphi$ (**from** (NF-Bel_i))
2. $\mathit{Bel}_i\, \mathit{After}_\alpha\, \mathit{Done}_\alpha\, \varphi \wedge \neg \mathit{Bel}_i\, \mathit{After}_\alpha\, \bot \rightarrow \mathit{After}_\alpha\, \mathit{Bel}_i\, \mathit{Done}_\alpha\, \varphi$ (**by** instantiation of 1.)
3. $\varphi \rightarrow \mathit{After}_\alpha\, \mathit{Done}_\alpha\, \varphi$ (**from** (CONV-AD))
4. $\mathit{Bel}_i\, \varphi \rightarrow \mathit{Bel}_i\, \mathit{After}_\alpha\, \mathit{Done}_\alpha\, \varphi$ (**from** 3. by (RM-$\Box$) for Bel_i)
5. $\mathit{Bel}_i\, \varphi \wedge \neg \mathit{Bel}_i\, \mathit{After}_\alpha\, \top \rightarrow \mathit{After}_\alpha\, \mathit{Bel}_i\, \mathit{Done}_\alpha\, \varphi$ (**from** 2. and 4. by $\mathcal{PL}$)
6. $\mathit{Done}_\alpha\, (\mathit{Bel}_i\, \varphi \wedge \neg \mathit{Bel}_i\, \mathit{After}_\alpha\, \top) \rightarrow \mathit{Bel}_i\, \mathit{Done}_\alpha\, \varphi$ (**by** Lemma 3)
7. $\mathit{Done}_\alpha\, \Phi \wedge \mathit{Done}_\alpha\, \Psi \rightarrow \mathit{Done}_\alpha\, (\Phi \wedge \Psi)$ (**from** (CD-DB))
8. $\mathit{Done}_\alpha\, \neg \mathit{Bel}_i\, \mathit{After}_\alpha\, \bot \wedge \mathit{Done}_\alpha\, \mathit{Bel}_i\, \varphi \rightarrow \mathit{Bel}_i\, \mathit{Done}_\alpha\, \varphi$ (**from** 6. and 7.) $\Box$

Proof 10 (of Theorem 9) *Case of (a).*

1. $\mathit{Fear}_i\, \neg \mathit{Happens}_{i:\alpha}\, \varphi \rightarrow \neg \mathit{Bel}_i\, \neg \mathit{Happens}_{i:\alpha}\, \varphi$ (**by** definition of Fear and definition 1)
2. $\mathit{Fear}_i\, \neg \mathit{Happens}_{i:\alpha}\, \varphi \rightarrow \neg \mathit{Bel}_i\, \neg \mathit{Happens}_{i:\alpha}\, \top$ (**from** 1. by (RM-$\Diamond$) for $\neg \mathit{Bel}_i\, \neg$)
3. $\mathit{Fear}_i\, \neg \mathit{Happens}_{i:\alpha}\, \varphi \rightarrow \neg \mathit{Bel}_i\, \mathit{After}_\alpha\, \bot$ (**from** 2. by definition of $\mathit{Happens}$)
4. $\vdash \mathit{Fear}_i\, \neg \mathit{Happens}_{i:\alpha}\, \varphi \wedge \mathit{Bel}_i\, \mathit{Idl}_i\, \mathit{Happens}_{i:\alpha}\, \varphi \rightarrow$ $\mathit{Bel}_i\, (\mathit{Fear}_i\, \neg \mathit{Happens}_{i:\alpha}\, \varphi \wedge \mathit{Idl}_i\, \mathit{Happens}_{i:\alpha}\, \varphi \wedge \neg \mathit{Bel}_i\, \mathit{After}_\alpha\, \bot)$ (**by** Theorem 13, (5-Bel_i) and (C-$\Box$) for Bel_i)
5. $\vdash \mathit{Fear}_i\, \neg \mathit{Happens}_{i:\alpha}\, \varphi \wedge \mathit{Bel}_i\, \mathit{Idl}_i\, \mathit{Happens}_{i:\alpha}\, \varphi \rightarrow \mathit{After}_{i:\alpha}\, \mathit{Done}_{i:\alpha}\, \mathit{Bel}_i$ $(\mathit{Fear}_i\, \neg \mathit{Happens}_{i:\alpha}\, \varphi \wedge \mathit{Idl}_i\, \mathit{Happens}_{i:\alpha}\, \varphi \neg \mathit{Bel}_i\, \mathit{After}_\alpha\, \bot)$ (**by** (CONV-AD))

6. $\vdash Fear_i\, \neg Happens_{i:\alpha}\, \varphi \wedge Bel_i\, Idl_i\, Happens_{i:\alpha}\, \varphi \rightarrow After_{i:\alpha}\, Bel_i\, Done_{i:\alpha}$
$(Fear_i\, \neg Happens_{i:\alpha}\, \varphi \wedge Idl_i\, Happens_{i:\alpha}\, \varphi)$ (**by** Lemma 2)

7. $\vdash Fear_i\, \neg Happens_{i:\alpha}\, \varphi \wedge Bel_i\, Idl_i\, Happens_{i:\alpha}\, \varphi \rightarrow After_{i:\alpha}\, (Bel_i\, \varphi \rightarrow$
$Bel_i\, Done_{i:\alpha}\, (Fear_i\, \neg Happens_{i:\alpha}\, \varphi \wedge Idl_i\, Happens_{i:\alpha}\, \varphi) \wedge Bel_i\, \varphi)$

8. $\vdash Fear_i\, \neg Happens_{i:\alpha}\, \varphi \wedge Bel_i\, Idl_i\, Happens_{i:\alpha}\, \varphi \rightarrow After_{i:\alpha}\, (Bel_i\, \varphi \rightarrow$
$Bel_i\, Done_{i:\alpha}\, (Prob_i\, After_{i:\alpha}\, \neg\varphi \wedge Idl_i\, Happens_{i:\alpha}\, \varphi) \wedge Bel_i\, \varphi)$
(**by** definitions of $Fear$ and $Happens$)

9. $\vdash Fear_i\, \neg Happens_{i:\alpha}\, \varphi \wedge Bel_i\, Idl_i\, Happens_{i:\alpha}\, \varphi \rightarrow$
$After_{i:\alpha}\, (Bel_i\, \varphi \rightarrow Pride_i\, (i{:}\alpha, \varphi))$ (**by** definition of $Pride$)

The proof is similar for (b), (c), and (d). □

Theorem 12 (Inconsistency between good-will and ill-will emotions)

$$\vdash \neg(HappyFor_{i,j}\varphi \wedge Resentment_{i,j}\varphi) \quad \text{(a)}$$
$$\vdash \neg(SorryFor_{i,j}\varphi \wedge Gloating_{i,j}\varphi) \quad \text{(b)}$$
$$\vdash \neg(HappyFor_{i,j}\varphi \wedge Gloating_{i,j}\varphi) \quad \text{(c)}$$
$$\vdash \neg(SorryFor_{i,j}\varphi \wedge Resentment_{i,j}\varphi) \quad \text{(d)}$$

Sketch of proof 2 (of Theorem 12) *The proof for cases (a) and (b) follows from the rationality of Des_i. The proof for cases (c) and (d) follows from Lemma 4.* □

Lemma 4 $\neg(Bel_i\, Des_j\, \varphi \wedge Bel_i\, Des_j\, \neg\varphi)$

Proof 11 (of Lemma 4)

1. $\vdash Des_j\, \varphi \rightarrow \neg Des_j\, \neg\varphi$ (**from** (D-Des_i))
2. $\vdash Bel_i\, Des_j\, \varphi \rightarrow Bel_i\, \neg Des_j\, \neg\varphi$ (**by** (RM-□) for Bel_i)
3. $\vdash Bel_i\, Des_j\, \varphi \rightarrow \neg Bel_i\, Des_j\, \neg\varphi$ (**by** (D-Bel_i))
4. $\vdash \neg(Bel_i\, Des_j\, \varphi \wedge Bel_i\, Des_j\, \neg\varphi)$ (**by** $\mathcal{PL}$) □

The face of emotions: a logical formalization of expressive speech acts

Nadine Guiraud
UPS, IRIT, France
Nadine.Guiraud@irit.fr

Dominique Longin
CNRS, IRIT, France
Dominique.Longin@irit.fr

Emiliano Lorini
CNRS, IRIT, France
Emiliano.Lorini@irit.fr

Sylvie Pesty
LIG, France
Sylvie.Pesty@imag.fr

Jérémy Rivière
LIG, France
jeremy.riviere@imag.fr

ABSTRACT

In this paper, we merge speech act theory, emotion theory, and logic. We propose a modal logic that integrates the concepts of belief, goal, ideal and responsibility and that allows to describe what a given agent expresses in the context of a conversation with another agent. We use the logic in order to provide a systematic analysis of expressive speech acts, that is, speech acts that are aimed at expressing a given emotion (e.g. to apologize, to thank, to reproach, etc.).

Categories and Subject Descriptors

I.2 [**Artificial Intelligence**]: Multiagent systems

General Terms

Theory

Keywords

Speech act theory, cognitive models, logic-based approaches and methods

1. INTRODUCTION

Since the works of Austin [2] and Searle [20] on speech acts, there has been a lot of work on illocutionary acts[1] and on their use for the formal specification of an agent communication language (see, e.g., [6, 27, 23, 9, 10]). Searle has defined five classes of illocutionary acts [21, Chapter 1], and every utterance realizes the performance of one (or more) illocutionary act(s) of theses classes. Thus, Searle's classification is a taxonomy. These fives classes of illocutionary acts are:

- assertives (for describing facts, *e.g.* "It rains"),
- directives (for representing order or request for instance, *e.g.* "Open the door, please"),
- commissives (for representing commitment, *e.g.* "I will help you"),
- declarations (for representing institutional illocutionary acts, *e.g.* "I name this ship the *Queen Elizabeth*"),
- expressives (for representing psychological attitudes, *e.g.* "I congratulate you" or "I thank you").

Existing literature on speech acts is mainly about the first three classes of illocutionary acts and, to a lesser extent, about the fourth. Thus, as far as we know, there is no work about the last class of illocutionary acts, that is, expressives. As Searle says:

> The illocutionary point of this class is to express the psychological state specified in the sincerity condition about a state of affairs specified in the propositional content. The paradigms of expressive verbs are "to thank", "to congratulate", "to apologize", "to deplore", and "to welcome". [21, Chapter 1]

In this paper we propose a first formalization of expressive speech acts in a BDI-like logic where utterances are represented by the mental states they express. The logic, which is presented in Section 2, has specific modal operators that allow us to represent *expressed* psychological mental states.

We focus on particular psychological states that are emotional states. Emotions that we consider are either *basic emotions* (only defined from beliefs and goals) or *complex emotions* (based on complex reasoning about norms, responsability, *etc.*). For instance, joy and sadness are basic emotions, whereas guilt or regret are complex emotions requiring a complex form counterfactual reasoning about responsibility where reality is compared to an imagined view of what might have been [12, 15]. Basic and complex emotions are studied in Section 3. In the paper we only consider the cognitive structure of emotion rather than emotion as a complex psychological phenomenon including cognitive aspects and somatic aspects (i.e. feeling). Indeed the cognitive structure of emotion is sufficient for our needs, as we only consider the mental states that can be expressed by use of language.

In Section 4, expressive speech acts are defined as public expressions of emotional states.

[1]Searle distinguishes several types of speech acts: utterance acts (using for uttering words); propositional acts (for referring and predicating); illocutionary acts (for stating, questioning, commanding, promising, *etc.*). See [20, Section 2.1] for more details.

Cite as: The face of emotions: a logical formalization of expressive speech acts, N. Guiraud, D. Longin, E. Lorini, S. Pesty and J. Rivière, *Proc. of 10th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2011)*, Tumer, Yolum, Sonenberg and Stone (eds.), May, 2–6, 2011, Taipei, Taiwan, pp. XXX-XXX.

2. LOGICAL FRAMEWORK

MLC (*Modal Logic of Communication*) is a BDI-like logic [7, 17] that allows us to represent agents' mental states (beliefs, desires and ideals) as well as the overt and social aspect of communication. It has modal operators that describe the conversational state of an agent i with respect to another agent j in front of an audience H, *i.e.* what agent i expresses to agent j in front of the audience H. A conversational state is a static description of the utterances that are performed by the participants in a dialogue, and is similar to the commitment store of Walton & Krabbe [28].

2.1 Syntax

Assume a finite non-empty set $AGT = \{1, \ldots, n\}$ of agents, a countable set $ATM = \{p, q, \ldots\}$ of atomic propositions denoting facts. The language $\mathcal{L}$ of the logic **MLC** is the set of formulas defined by the following BNF:

$$\varphi \ ::= \ p \ \mid \ \neg\varphi \ \mid \ \varphi \wedge \varphi \ \mid \ \mathbf{Bel}_i\,\varphi \ \mid \ \mathbf{Goal}_i\varphi \ \mid \ \mathbf{Ideal}_i\varphi \ \mid \ \mathbf{Cd}_i\varphi \ \mid \ \mathbf{Exp}_{i,j,H}\varphi$$

where p ranges over ATM, i, j range over AGT and H ranges over 2^{AGT}. The other Boolean constructions $\top$, $\bot$, $\vee$, $\rightarrow$ and $\leftrightarrow$ are defined in the standard way.

Operators $\mathbf{Bel}_i$ and $\mathbf{Goal}_i$ are used to represent agent i's beliefs and goals. Given an arbitrary formula φ of the logic, $\mathbf{Bel}_i\,\varphi$ has to be read 'agent i believes that φ', whereas $\mathbf{Goal}_i\varphi$ has to be read 'agent i has the goal that φ' or 'agent i wants φ to be true'. Following [8], we consider goals the most basic class of motivational attitudes. The concept of goal is more general than the concept of desire (therefore, the former class includes the latter). Desires are intrinsically endogenous, while goals might originate from external inputs.[2] For instance, an agent might have a goal because of norm compliance or because it adopted this goal from another agent (*e.g.* agent i has the goal to close the door because agent j asked it to do so and i accepted j's request). Moreover, differently from a desire, a goal is not necessarily associated with a pleasant state of mind (*i.e.* goals do not necessarily have a hedonistic component).

As the class of goals includes desires, we assume that goals can be incompatible with beliefs. For instance, a person may wish to become multimillionaire even though she believes that her aspiration will never be satisfied.

The operators $\mathbf{Ideal}_i$ are used to represent an agent's moral attitudes, after supposing that agents are capable to discern what (from their point of view) is morally right from what is morally wrong. This is a necessary step towards an analysis of social emotions such as guilt and shame which involve a moral dimension. The formula $\mathbf{Ideal}_i\varphi$ means 'φ is an ideal state of affairs for agent i'. More generally, $\mathbf{Ideal}_i\varphi$ expresses that agent i thinks that it ought to promote the realization of φ, that is, agent i conceives a demanding connection between itself and the state of affairs φ. When agent i endorses the ideal that φ (*i.e.* $\mathbf{Ideal}_i\varphi$ is true), it means that i addresses a command to itself, or a request or an imperative to achieve φ (when φ is actually false) or to maintain φ (when φ is actually true) [4]. In this sense, i feels morally responsible for the realization of φ.

There are different ways to explain how a state of affairs φ becomes an ideal state of affairs of an agent. A plausible explanation is based on the hypothesis that ideals are just social norms internalized (or adopted) by an agent (see [8] for a general theory of norm internalization). Suppose that an agent believes that in a certain group (or institution) there exists a certain norm (e.g. an obligation) prescribing that a state of affairs φ should be true. Moreover, assume that the agent identifies itself as a member of this group. In this case, the agent adopts the norm, that is, the external norm becomes an ideal of the agent. For example, since I believe that in Italy it is obligatory to pay taxes and I identify myself as an Italian citizen, I adopt this obligation by imposing the imperative to pay taxes to myself.

The operators $\mathbf{Cd}_i$ are used to talk about agents' choices and actions, and will be later used in order to define a basic notion of responsibility. Formula $\mathbf{Cd}_i\varphi$ has to be read 'given what the other agents have done, agent i could have ensured φ to be true' or 'given what the other agents have decided to do, agent i could have ensured φ to be true'. Similar operators have been studied in [15] in the framework of STIT logic (the logic of *Seeing to it that*) [11] in order to provide an analysis of counterfactual emotions such as regret and disappointment.

Finally, formula $\mathbf{Exp}_{i,j,H}\varphi$ has to be read 'agent i expressed to agent j that φ is true in front of group H'. Given a formula $\mathbf{Exp}_{i,j,H}\varphi$, we call i the *speaker*, j the *addressee*, H the *audience* and φ the *content* of the speaker's expression. For example, we can represent the sentence "John told to Mary: I have a new car." by the formula $\mathbf{Exp}_{John,Mary,H}\ newCar$ where H are the agents who can hear John's speech act. The basic function of modalities $\mathbf{Exp}_{i,j,H}$ is to keep trace of the information that agent i has communicated to agent j in front of an audience H.

Further concepts.

We define a basic concept of responsibility as follows:

$$\mathbf{Resp}_i\varphi \stackrel{def}{=} \varphi \wedge \mathbf{Cd}_i\neg\varphi$$

According to this definition, 'agent i is responsible for φ' (noted $\mathbf{Resp}_i\varphi$) if and only if, 'φ is true and, given what the other agents have done, i could have ensured φ to be false' which is the same thing as saying 'φ is true and i could have prevented φ to be true'. In other words, agent i is responsible for φ only if, there is a counterfactual dependence between the state of affairs φ and agent i's choice.[3] The concept of inevitability is defined as the dual of the operator $\mathbf{Cd}_i$:

$$\mathbf{Inev}_i\varphi \stackrel{def}{=} \neg\mathbf{Cd}_i\neg\varphi$$

Thus, 'φ is inevitable for agent i' (noted $\mathbf{Inev}_i\varphi$) if and only if, it is not the case that, given what the other agents have done, i could have ensured φ to be false.

We define one more concepts which will be useful for the analysis of expressive speech acts such as *to sympathize*, *to apologize* and *to be sorry for* proposed in Section 4. We say that 'agent i is willing to adopt agent j's goal that φ' or 'agent i is cooperative about φ with regard to agent j' (noted $\mathbf{AdoptGoal}_{i,j}\varphi$) if and only if, if i believes that j

[2] See [22] for a detailed analysis of how an agent may want something without desiring it and on the problem of *reasons for acting* independent from desires.

[3] This view of responsibility is close to that of [15, 5]. A stronger view of responsibility requires that agent i is responsible for φ only if it brings about φ, no matter what the other agents do.

wants φ to be true then i too wants φ to be true:[4]

$$\mathbf{AdoptGoal}_{i,j}\varphi \stackrel{def}{=} \mathbf{Bel}_i\,\mathbf{Goal}_j\varphi \rightarrow \mathbf{Goal}_i\varphi$$

2.2 Semantics

We use a standard possible worlds semantics where accessibility relations are used to interprete the modal operators of our logic. **MLC**-models are tuples $M = \langle W, \mathcal{B}, \mathcal{G}, \mathcal{I}, \mathcal{O}, \mathcal{E}, \mathcal{V}\rangle$ defined as follows:

- W is a nonempty set of possible *worlds* or *states*;
- $\mathcal{B} : AGT \longrightarrow 2^{W\times W}$ maps every agent $i \in AGT$ to a serial,[5] transitive[6] and Euclidean[7] relation $\mathcal{B}_i$ over W;
- $\mathcal{G} : AGT \longrightarrow 2^{W\times W}$ maps every agent $i \in AGT$ to a serial relation $\mathcal{G}_i$ over W;
- $\mathcal{I} : AGT \longrightarrow 2^{W\times W}$ maps every agent $i \in AGT$ to a serial relation $\mathcal{I}_i$ over W;
- $\mathcal{O} : AGT \longrightarrow 2^{W\times W}$ maps every agent $i \in AGT$ to an equivalence (*i.e.* reflexive,[8] transitive and symmetric[9]) relation $\mathcal{O}_i$ over W;
- $\mathcal{E} : AGT \times AGT \times 2^{AGT} \longrightarrow 2^{W\times W}$ maps every pair of agents $i, j \in AGT$ and set of agents $H \in 2^{AGT}$ to a transitive relation $\mathcal{E}_{i,j,H}$ over W;
- $\mathcal{V} : ATM \longrightarrow 2^W$ is a valuation function.

Moreover, we write $\mathcal{B}_i(w) = \{v|(w,v) \in \mathcal{B}_i\}$, $\mathcal{G}_i(w) = \{v|(w,v) \in \mathcal{G}_i\}$, $\mathcal{I}_i(w) = \{v|(w,v) \in \mathcal{I}_i\}$, $\mathcal{O}_i(w) = \{v|(w,v) \in \mathcal{O}_i\}$ and $\mathcal{E}_{i,j,H}(w) = \{v|(w,v) \in \mathcal{E}_{i,j,H}\}$.

The set $\mathcal{B}_i(w)$ is the *information state* of agent i at world w: the set of worlds that agent i considers possible at world w. The fact that every $\mathcal{B}_i$ is serial means that an agent has always consistent beliefs. Moreover, the transitivity and Euclideanity of $\mathcal{B}_i$ mean that an agent's beliefs are positively and negatively introspective.

The set $\mathcal{G}_i(w)$ is the *goal state* of agent i at world w: the set of worlds that agent i wants to reach (or prefers) at world w. The fact that every $\mathcal{G}_i$ is serial means that an agent has always at least one state that it wants to reach.

The set $\mathcal{I}_i(w)$ is the *ideal state* of agent i at world w: the set of worlds that agent i considers ideal (from a moral point of view) at world w. The fact that every $\mathcal{I}_i$ is serial means that an agent has always at least one ideal state.

The set $\mathcal{O}_i(w)$ is the *outcome state* of agent i at world w: $\mathcal{O}_i(w)$ is the set of outcomes that agent i could have ensured at w, given what the other agents have done (at w). Therefore, the fact that $\mathcal{O}_i$ is reflexive means that the actual world is an outcome that agent i could have ensured, given what the other agents have done. The fact that $\mathcal{O}_i$ is transitive means if v is an outcome that agent i can ensure at w and u is an outcome that agent i can ensure at v then u is an outcome that agent i can ensure at w. The fact that $\mathcal{O}_i$ is Euclidean means if v is an outcome that agent i can ensure at w and u is an outcome that agent i can ensure at w then u is an outcome that agent i can ensure at v.

Finally, the set $\mathcal{E}_{i,j,H}(w)$ is the *conversational state* of agent i with respect to agent j in the presence of group H at world w: the set of worlds that are compatible with what has been expressed by agent i to agent j in front of group H at world w. The fact that $\mathcal{E}_{i,j,H}$ is transitive means that if v is compatible with what has been expressed by agent i to agent j in front of group H at w and u is compatible with what has been expressed by agent i to agent j in front of group H at v, then if u is compatible with what has been expressed by agent i to agent j in front of group H at w. Note that $\mathcal{E}_{i,j,H}(w)$ is different from $\mathcal{B}_i(w)$ because what agent i has expressed may be different from what agent i believes (case of insincerity).

MLC-models are supposed to satisfy the following additional constraints. For every world $w \in W$, for all $i, j, z \in AGT$, for all $H \in 2^{AGT}$, if $z \in H \cup \{i, j\}$ then:

S1 if $v \in \mathcal{B}_i(w)$ then $\mathcal{G}_i(v) = \mathcal{G}_i(w)$;

S2 if $v \in \mathcal{B}_i(w)$ then $\mathcal{I}_i(v) = \mathcal{I}_i(w)$;

S3 if $v \in \mathcal{B}_z(w)$ then $\mathcal{E}_{i,j,H}(v) = \mathcal{E}_{i,j,H}(w)$.

Constraint S1 is a property of positive and negative introspection for goals: worlds that are preferred by agent i are also preferred by agent i from those worlds that it considers possible. Constraint S2 is the corresponding property of positive and negative introspection for ideals. Constraint S3 is a property of positive and negative introspection for communication. Suppose that $z \in H \cup \{i, j\}$. Then, S3 means that: worlds that are compatible with what agent i expressed to agent j in front of group H, are also compatible with what agent i expressed to agent j in front of group H from those worlds that agent z considers possible.

Given a model M, a world w and a formula φ, we write $M, w \models \varphi$ to mean that φ is true at world w in M. Truth conditions of formulas are defined as follows:

- $M, w \models p$ iff $w \in \mathcal{V}(p)$;
- $M, w \models \neg\varphi$ iff not $M, w \models \varphi$;
- $M, w \models \varphi \wedge \psi$ iff $M, w \models \varphi$ and $M, w \models \psi$;
- $M, w \models \mathbf{Bel}_i\,\varphi$ iff $M, v \models \varphi$ for all $v \in \mathcal{B}_i(w)$;
- $M, w \models \mathbf{Goal}_i\varphi$ iff $M, v \models \varphi$ for all $v \in \mathcal{G}_i(w)$;
- $M, w \models \mathbf{Ideal}_i\varphi$ iff $M, v \models \varphi$ for all $v \in \mathcal{I}_i(w)$;
- $M, w \models \mathbf{Cd}_i\varphi$ iff $M, v \models \varphi$ for some $v \in \mathcal{O}_i(w)$;
- $M, w \models \mathbf{Exp}_{i,j,H}\varphi$ iff $M, v \models \varphi$ for all $v \in \mathcal{E}_{i,j,H}(w)$.

Note that while the operators $\mathbf{Bel}_i$, $\mathbf{Goal}_i$, $\mathbf{Ideal}_i$ and $\mathbf{Exp}_{i,j,H}$ are all $\Box$ ('Box') modal operators, $\mathbf{Cd}_i$ are $\Diamond$ ('Diamond') modal operators. That is, an agent i could have ensured φ at w of world M (*i.e.* $M, w \models \mathbf{Cd}_i\varphi$) if and only if there is an outcome that agent i can ensure at w, given what the other agents have done (at w), in which φ is true.

As usual we say that φ is *valid* in **MLC** (noted $\models_{\mathbf{MLC}} \varphi$) iff for all models $M = \langle W, \mathcal{B}, \mathcal{G}, \mathcal{I}, \mathcal{O}, \mathcal{E}, \mathcal{V}\rangle$ and for all worlds $w \in W$ we have $M, w \models \varphi$.

[4] We are aware that some form of conditional rather than material implication would be more suited to express entailment in the notion of goal adoption.

[5] A given relation $\mathcal{R}$ on W is serial if and only if for every $w \in W$ there is v such that $(w, v) \in \mathcal{R}$.

[6] A given relation $\mathcal{R}$ on W is transitive if and only if, if $(w, v) \in \mathcal{R}$ and $(v, u) \in \mathcal{R}$ then $(w, u) \in \mathcal{R}$.

[7] A given relation $\mathcal{R}$ on W is Euclidean if and only if, if $(w, v) \in \mathcal{R}$ and $(w, u) \in \mathcal{R}$ then $(v, u) \in \mathcal{R}$.

[8] A given relation $\mathcal{R}$ on W is reflexive if and only if for every $w \in W$, $(w, w) \in \mathcal{R}$.

[9] A given relation $\mathcal{R}$ on W is symmetric if and only if, if $(w, v) \in \mathcal{R}$ then $(v, w) \in \mathcal{R}$.

2.3 Axiomatization

All KD45-principles for the operators $\mathbf{Bel}_i$	($\mathrm{KD45}_{\mathbf{Bel}}$)
All KD-principles for the operators $\mathbf{Goal}_i$	($\mathrm{KD}_{\mathbf{Goal}}$)
All KD-principles for the operators $\mathbf{Ideal}_i$	($\mathrm{KD}_{\mathbf{Ideal}}$)
All S5-principles for the operators $\mathbf{Cd}_i$	($\mathrm{S5}_{\mathbf{Cd}}$)
All K4-principles for the operators $\mathbf{Exp}_{i,j,H}$	($\mathrm{K4}_{\mathbf{Express}}$)
$\mathbf{Goal}_i\varphi \rightarrow \mathbf{Bel}_i\,\mathbf{Goal}_i\varphi$	($\mathrm{PI}_{\mathbf{Goal}}$)
$\neg\mathbf{Goal}_i\varphi \rightarrow \mathbf{Bel}_i\,\neg\mathbf{Goal}_i\varphi$	($\mathrm{NI}_{\mathbf{Goal}}$)
$\mathbf{Ideal}_i\varphi \rightarrow \mathbf{Bel}_i\,\mathbf{Ideal}_i\varphi$	($\mathrm{PI}_{\mathbf{Ideal}}$)
$\neg\mathbf{Ideal}_i\varphi \rightarrow \mathbf{Bel}_i\,\neg\mathbf{Ideal}_i\varphi$	($\mathrm{NI}_{\mathbf{Ideal}}$)
$\mathbf{Exp}_{i,j,H}\varphi \rightarrow \mathbf{Bel}_z\,\mathbf{Exp}_{i,j,H}\varphi$ (if $z \in H \cup \{i,j\}$)	($\mathrm{PI}_{\mathbf{Express}}$)
$\neg\mathbf{Exp}_{i,j,H}\varphi \rightarrow \mathbf{Bel}_z\,\neg\mathbf{Exp}_{i,j,H}\varphi$ (if $z \in H \cup \{i,j\}$)	($\mathrm{NI}_{\mathbf{Express}}$)

Figure 1: Axiomatization of MLC

Figure 1 contains the axiomatization of the logic **MLC**. We have all principles of the normal modal logic KD45 for every belief operator $\mathbf{Bel}_i$. Thus, an agent cannot have inconsistent beliefs (*i.e.* $\neg(\mathbf{Bel}_i\,\varphi \wedge \mathbf{Bel}_i\,\neg\varphi)$), and it has positive and negative introspection over its beliefs (*i.e.* $\mathbf{Bel}_i\,\varphi \rightarrow \mathbf{Bel}_i\,\mathbf{Bel}_i\,\varphi$ and $\neg\mathbf{Bel}_i\,\varphi \rightarrow \mathbf{Bel}_i\,\neg\mathbf{Bel}_i\,\varphi$).

We have all principles of the normal modal logic KD for every operator $\mathbf{Goal}_i$ and for every operator $\mathbf{Ideal}_i$ (*i.e.* $\neg(\mathbf{Goal}_i\varphi \wedge \mathbf{Goal}_i\neg\varphi)$ and $\neg(\mathbf{Ideal}_i\varphi \wedge \mathbf{Ideal}_i\neg\varphi)$).

We have all principles of the normal modal logic S5 for every operator $\mathbf{Cd}_i$, taking it as a 'Diamond' operator. Thus, for example, if φ is true then an agent could have ensured φ (*i.e.* $\varphi \rightarrow \mathbf{Cd}_i\varphi$).

Moreover, we have all principles of the normal modal logic K4 for every communication operator $\mathbf{Exp}_{i,j,H}$. Thus, i's action of expressing to j that φ entails i's action of expressing to j that i expresses to j that φ (*i.e.* $\mathbf{Exp}_{i,j,H}\varphi \rightarrow \mathbf{Exp}_{i,j,H}\mathbf{Exp}_{i,j,H}\varphi$). In other words, the action of expressing something to someone has a self-referential nature. We do not include Axiom D for the operator $\mathbf{Exp}_{i,j,H}$. Thus, we accept that an agent may express inconsistent things to another agent (even though it cannot believe them), that is, we accept formula $\mathbf{Exp}_{i,j,H}\bot$ to be satisfiable in our logic.

Axioms ($\mathrm{PI}_{\mathbf{Goal}}$) and ($\mathrm{NI}_{\mathbf{Goal}}$) are standard axioms of positive and negative introspection for goals [14], while Axioms ($\mathrm{PI}_{\mathbf{Ideal}}$) and ($\mathrm{NI}_{\mathbf{Ideal}}$) are corresponding principles for ideals.

Finally, Axioms ($\mathrm{PI}_{\mathbf{Express}}$) and ($\mathrm{NI}_{\mathbf{Express}}$) are corresponding principles of positive and negative introspection for communication: if an agent i expressed (resp. did not express) something to another agent j in front of an audience H, then this is public for the group $H \cup \{i,j\}$ including the speaker, the addressee, and all agents in the audience.

Note that we did not include a general inclusion principle of the form:

$$\mathbf{Exp}_{i,j,H}\varphi \rightarrow \mathbf{Exp}_{i,j,I}\varphi \text{ for } I \subseteq H$$

In fact, we want to be able to model situations in which an agent i expressed something in secret to another agent j (while all other agents were not hearing), and it expressed the contrary to j in front of a larger group including j, without expressing an inconsistency.

For example, Bill might express in secret to Mary that he loves Ann, *i.e.* $\mathbf{Exp}_{Bill,Mary,\emptyset}\mathit{BillLovesAnn}$, and express to Mary that he does not love Ann when he is in front of Bob, *i.e.* $\mathbf{Exp}_{Bill,Mary,\{Bob\}}\neg\mathit{BillLovesAnn}$, without expressing an inconsistency in front of Mary, *i.e.* $\neg\mathbf{Exp}_{Bill,Mary,\emptyset}\bot$.

THEOREM 1. *The axiomatization in Figure 1 is sound and complete with respect to the class of **MLC**-models.*

PROOF (SKETCH). It is a routine task to check that the axioms of the logic **MLC** correspond one-to-one to their semantic counterparts on the models.

In particular, ($\mathrm{KD45}_{\mathbf{Bel}}$) corresponds to the fact that every $\mathcal{B}_i$ is serial, transitive and Euclidean. ($\mathrm{KD}_{\mathbf{Goal}}$) and ($\mathrm{KD}_{\mathbf{Ideal}}$) correspond to the fact that every $\mathcal{G}_i$ (resp. $\mathcal{I}_i$) is serial. ($\mathrm{S5}_{\mathbf{Cd}}$) corresponds to the fact that every $\mathcal{O}_i$ is an equivalence relation, while ($\mathrm{K4}_{\mathbf{Express}}$) corresponds to the transitivity of every $\mathcal{E}_{i,j,H}$. Axioms ($\mathrm{PI}_{\mathbf{Goal}}$) and ($\mathrm{NI}_{\mathbf{Goal}}$) together correspond to the Constraint S1, Axioms ($\mathrm{PI}_{\mathbf{Ideal}}$) and ($\mathrm{NI}_{\mathbf{Ideal}}$) together correspond to the Constraint S2. Axioms ($\mathrm{PI}_{\mathbf{Express}}$) and ($\mathrm{NI}_{\mathbf{Express}}$) together correspond to the Constraint S3. It is routine, too, to check that all axioms of the logic **MLC** are in the Sahlqvist class. This means that the axioms are all expressible as first-order conditions on models and that they are complete with respect to the defined model classes, cf. [3, Th. 2.42]. □

We write $\vdash_{\mathbf{MLC}}\varphi$ if φ is a **MLC**-theorem. The following are examples of **MLC**-theorems. For every $i,j \in AGT$ and for every $H \in 2^{AGT}$ we have:

$$\vdash_{\mathbf{MLC}} \mathbf{Exp}_{i,j,H}\varphi \leftrightarrow \bigwedge_{z \in H\cup\{i,j\}} \mathbf{Bel}_z\,\mathbf{Exp}_{i,j,H}\varphi$$

$$\vdash_{\mathbf{MLC}} \neg\mathbf{Exp}_{i,j,H}\varphi \leftrightarrow \bigwedge_{z \in H\cup\{i,j\}} \mathbf{Bel}_z\,\neg\mathbf{Exp}_{i,j,H}\varphi$$

According to former formula, agent i has expressed that φ to j in front of the audience H if and only if, i,j and every agent in the audience believes this. According to the latter, agent i did not express that φ to j in front of the audience H if and only if i,j and every agent in the audience believes this.

3. FORMALIZATION OF EMOTIONS

As said in Section 1, Searle says that expressives are expressions of psychological states. Vanderveken agree with this and says that such psychological states have the logical form $m(p)$ where m is the psychological mode and p "the propositional content which represents the state of affairs to which [the act is] directed" [26, p. 213]. Here, emotions are viewed as particular mental states that have the logical form $m(p)$. Thus, emotion is here always about a state of affairs. When it is not the case, we consider such feeling to be a mood rather than an emotion. We are not concerned here by mood.

Following dimensional theories of emotion [18], the difference between two close labels in a multi-dimensional space may be a difference of intensity of the same emotion. It means that their cognitive structure is the same. In this paper we do not deal with intensity of emotions and we only formalize cognitive structures of emotions rather than

emotions themselves. Following appraisal theories [19, 13], the cognitive structure of an emotion is the configuration of mental states that an agent has in mind when feeling this emotion and that is responsible for this feeling. It is just a part of the entire affective phenomenon.

In the rest of this article, we use the term *emotion* to refer to the *cognitive structure of emotion*. The definitions of emotions will be written in italic in order to distinguish them from the definitions of expressive speech acts given in Section 4.

3.1 Cognitive structure of basic emotions

Basic emotions concern emotions built from belief, and goals or ideals. When agent i believes that φ is true, if it aims at φ then it feels joy about the fact that φ is true; if it aims at $\neg\varphi$ then it feels sadness about the fact that φ is true; if it thinks that φ is an ideal state of affairs then it feels approval; finally, if it thinks that $\neg\varphi$ is an ideal state of affairs then it feels disapproval. These emotions are summarized in the following table.

$\wedge$	$\mathbf{Goal}_i\varphi$	$\mathbf{Goal}_i\neg\varphi$	$\mathbf{Ideal}_i\varphi$	$\mathbf{Ideal}_i\neg\varphi$
$\mathbf{Bel}_i\,\varphi$	$Joy_i\,\varphi$	$Sadness_i\,\varphi$	$Approval_i\varphi$	$Disapproval_i\varphi$

Agent i feels joy about φ if and only if, i believes that φ is true and wants φ to be true:

$$Joy_i\,\varphi \stackrel{def}{=} \mathbf{Bel}_i\,\varphi \wedge \mathbf{Goal}_i\varphi$$

For example, agent i feels joy for having passed the exam because i believes that it has passed the exam and wants to pass the exam. In this sense, i is pleased by the fact that it believes to have achieved what it wanted to achieve. This means that *joy* has a positive valence, that is, it is associated with goal achievement.[10]

Consider now sadness:

$$Sadness_i\,\varphi \stackrel{def}{=} \mathbf{Bel}_i\,\varphi \wedge \mathbf{Goal}_i\neg\varphi$$

That is, agent i feels sadness about φ if and only if i believes that φ is true and wants $\neg\varphi$ to be true. For instance, agent i feels sad for not having passed the exam because i believes that it has not passed the exam and wants to pass the exam. In this sense, i is displeased by the fact that it believes not to have achieved what it was committed to achieve. This means that sadness has a negative valence, that is, it is associated with goal frustration.

When φ concerns ideals, agent i approves φ or i disapproves φ, depending respectively on the fact that φ is ideal or not ideal for it. Thus:

$$Approval_i\varphi \stackrel{def}{=} \mathbf{Bel}_i\,\varphi \wedge \mathbf{Ideal}_i\varphi$$

$$Disapproval_i\varphi \stackrel{def}{=} \mathbf{Bel}_i\,\varphi \wedge \mathbf{Ideal}_i\neg\varphi$$

Note that we refer here to the expressive part of approval and of disapproval. In fact, approval and disapproval are both expressives and declarations in Speech Act theory. There also exists a normative sense (like in: The judge says "I disapprove your release on parole [and thus, you come back to the jail]") that corresponds to a declaration in accordance with law (and not necessary with the internal psychological state of the judge). Here we focus on the expressive sense.

[10]The terms positive valence and negative valence are used by Ortony et al. [16], whereas Lazarus [13] uses the terms goal congruent *versus* goal incongruent emotions.

3.2 Cognitive structure of complex emotions

As said in the introduction, the cognitive structures of complex emotions include complex reasoning about norms, responsibility, *etc.* In the following, we suppose that agent i feels an emotion related to its own responsibility or related to the responsibility of agent j (supposed to be different from agent i) about φ. At the same time, when φ (respectively $\neg\varphi$) is a goal or an ideal of agent i, thus we can expect that agent i feels an emotion about φ.

There are many psychological models of emotions in the literature. One of the most widely accepted model in AI is that of Ortony, Clore and Collins [16], which defines emotions such as reproach, shame and anger that have already been formalized in logic (e.g. [1, 24]). However this model does not define emotions such as guilt or regret that are based on the concept of responsibility about actions and choices. Indeed, several psychologists (e.g. [13]) showed that guilt involves the conviction of having injured someone or of having violated some norm or imperative, and the belief that this could have been avoided. Similarly, many psychologists (e.g. [29, 12]) agree in considering regret as a negative, cognitively determined emotion that we experience when realizing or imagining that our present situation would have been better, had we acted differently. Our formalization of complex emotions such as regret and guilt follows this latter work in the area of psychology of emotions. (See also [15] a logical formalization of regret and [25] for a logical formalization of guilt.)

For instance, when agent i believes that it is responsible for φ while it has $\neg\varphi$ as a goal, agent i feels regret, and *vice versa*. Formally:

$$Regret_i\varphi \stackrel{def}{=} \mathbf{Goal}_i\neg\varphi \wedge \mathbf{Bel}_i\,\mathbf{Resp}_i\varphi$$

Imagine a situation in which there are only two agents i and j, that is, $AGT = \{i, j\}$. Agent i decides to park its car in a no parking area. Agent j (the policeman) fines agent i 100 €. Agent i regrets for having been fined 100 € (noted $Regret_i$*fine*). This means that, i wants not to be fined (noted $\mathbf{Goal}_i\neg$*fine*) and believes that it is responsible for having been fined (noted $\mathbf{Bel}_i\,\mathbf{Resp}_i$*fine*). That is, agent i believes that it has been fined 100 € and believes that it could have avoided to be fined (by parking elsewhere).

As $\mathbf{Bel}_i\,\mathbf{Resp}_i\varphi \rightarrow \mathbf{Bel}_i\,\varphi$, we have the following theorem.

THEOREM 2.

$$Regret_i\varphi \rightarrow Sadness_i\,\varphi$$

This means that if agent i regrets for φ, then it feels sad about φ. In the previous example, agent i regrets for having been fined 100 € which entails that it is sad for having been fined 100 €.

When agent i believes that agent j is responsible for φ, and i has $\neg\varphi$ as a goal, i is disappointed about φ. Formally:

$$Disappointment_{i,j}\varphi \stackrel{def}{=} \mathbf{Goal}_i\neg\varphi \wedge \mathbf{Bel}_i\,\mathbf{Resp}_j\varphi$$

Note that disappointment may have different degrees of intensity. Thus, a strong disappointment is closer to anger.

In a similar way, agent i feels guilty for φ (noted $Guilt_i\varphi$) if and only if $\neg\varphi$ is an ideal state of affairs for i (noted $\mathbf{Ideal}_i\neg\varphi$) and i believes that it is responsible for φ. Formally:

$$Guilt_i\varphi \stackrel{def}{=} \mathbf{Ideal}_i\neg\varphi \wedge \mathbf{Bel}_i\,\mathbf{Resp}_i\varphi$$

Thus, *regret* concerns goals whereas *guilt* concerns ideals. For example, imagine a situation in which there are only two agents i and j (that is $AGT = \{i, j\}$). Agent i decides to shoot with a gun and accidentally kills agent j. Agent i feels guilty for having killed someone (noted $Guilt_i killedSomeone$). This means that, i addresses an imperative to itself not to kill other people (noted $\mathbf{Ideal}_i \neg killedSomeone$) and agent i believes that it is responsible for having killed someone (noted $\mathbf{Resp}_i killedSomeone$).

We do not give more details about the cognitive structure of complex emotions. All these emotions are summarized in the following table:

$\wedge$	$\mathbf{Bel}_i\,\mathbf{Resp}_i\varphi$	$\mathbf{Bel}_i\,\mathbf{Resp}_j\varphi$
$\mathbf{Goal}_i\varphi$	$Rejoicing_i\varphi$	$Gratitude_{i,j}\varphi$
$\mathbf{Goal}_i\neg\varphi$	$Regret_i\varphi$	$Disappointment_{i,j}\varphi$
$\mathbf{Ideal}_i\varphi$	$MoralSatisfaction_i\varphi$	$Admiration_{i,j}\varphi$
$\mathbf{Ideal}_i\neg\varphi$	$Guilt_i\varphi$	$Reproach_{i,j}\varphi$

4. EXPRESSIVE SPEECH ACTS

As Searle says [20, Section 3.4]: "Wherever there is a psychological state specified in the sincerity condition, the performance of the act counts as an *expression* of that psychological state.This law holds whether the act is sincere or insincere, that is whether the speaker actually has the specified psychological state or not. (...) To thank, welcome or congratulate counts as *an expression of gratitude, pleasure* (at H's arrival) or *pleasure* (at H's good fortune)".[11] This is true for every class of illocutionary acts not only for expressives.

The sincerity condition of expressives is that the speaker has the psychological states that he/she expresses when he/she performs an expressive act. In others words, when agent i congratulates agent j about some φ related to j, the sincerity condition is that i is pleased about φ. "To congratulate" is nothing but the expression of its sincerity condition [20, Section 3.4].

Formally, if we note $\mu(\varphi)$ an emotion about the proposition φ, we characterize the performance of an expressive as the expression of $\mu(\varphi)$ from a speaker i to an addressee j in front of a group of agents H as follows: $\mathbf{Exp}_{i,j,H}\mu(\varphi)$.

Note that the expression of a proposition (of the form $\mathbf{Exp}_{i,j,H}\varphi$) and the expressive (of the form $\mathbf{Exp}_{i,j,H}\mu(\varphi)$) should not be mixed up: an expressive is the expression of a particular proposition (that is, a psychological state, an emotion) but the expression of a proposition is not necessarily an expressive. For instance, we can express a commitment and the corresponding illocutionary act is a commissive; or we can express our intention that the speaker does something, and the corresponding act is a directive.[12]

When every action is publicly performed, H represents the set of all agents AGT. In this case, if an agent says something, everybody knows that. The parameter H in the formula $\mathbf{Exp}_{i,j,H}\mu(\varphi)$ becomes useful in case of a private conversation within a group, where illocutionary acts are not publicly performed. For instance, suppose that a group of friends are together at a party. Suppose also that John is sad because he lost his cat. He wants to share his sorrow with Beth but not with the rest of the group. In this case, H is reduced to the empty set. Thus, the formula characterizing this situation is: $\mathbf{Exp}_{John,Beth,\emptyset} Sadness_{John}\, catDeath$.

[11] H stands for the hearer.

[12] Thus, our language enables to formalize classes of speech acts that we do not use here. As explained in Section 5, formalization of others classes will be done in future work. Here, we focus on expressives because this class of illocutionary acts is the least studied in literature.

4.1 Expression of basic emotions

We propose to represent expressive speech acts as particular assertive speech acts where the propositional content is about a psychological state. More precisely, it is the emotion that the speaker wants to express. For instance, when agent i wants to express to agent j its joy about φ (we call this act: to be delighted about φ), i asserts to j that it feels joy about the fact that φ is true. In the same way, to express sadness about the fact that φ is true, it is to be saddened by the fact that φ is true. In the expressive sense, *to express his/her (dis)approval* is *to (dis)approve of*. Thus, formally:

$$\mathbf{IsDelighted}_{i,j,H}\varphi \stackrel{def}{=} \mathbf{Exp}_{i,j,H}\, Joy_i\, \varphi$$

$$\mathbf{IsSaddened}_{i,j,H}\varphi \stackrel{def}{=} \mathbf{Exp}_{i,j,H}\, Sadness_i\, \varphi$$

$$\mathbf{ApprovesOf}_{i,j,H}\varphi \stackrel{def}{=} \mathbf{Exp}_{i,j,H}\, Approval_i\varphi$$

$$\mathbf{DisapprovesOf}_{i,j,H}\varphi \stackrel{def}{=} \mathbf{Exp}_{i,j,H}\, Disapproval_i\varphi$$

Note that in the case of disapproval, and following Vanderveken [26, p. 216], "it is not presupposed that the hearer is responsible for the state of affairs". Thus, we do not necessarily have that agent j is responsible for φ.

We say that agent i expresses to agent j that it is sorry for φ if and only if, i expresses to agent j that it is sad about the fact that j did not achieve its goal that $\neg\varphi$ (*i.e.* agent j has $\neg\varphi$ as a goal and φ is true):

$$\begin{aligned}\mathbf{IsSorryFor}_{i,j,H}\varphi &\stackrel{def}{=} \mathbf{IsSaddened}_{i,j,H}(\mathbf{Goal}_j\neg\varphi \wedge \varphi)\\ &\stackrel{def}{=} \mathbf{Exp}_{i,j,H}\, Sadness_i\, (\mathbf{Goal}_j\neg\varphi \wedge \varphi)\end{aligned}$$

The expressive *to sympathize* adds to the expressive *to be sorry for* an aspect of goal adoption. More precisely, agent i sympathizes with agent j for the fact that φ is true if and only if, i expresses sadness about the fact that agent j did not achieve its goal that $\neg\varphi$ (*i.e.* i expresses to j that it is sorry for φ) and i expresses that it is willing to adopt j's goal that $\neg\varphi$:

$$\begin{aligned}\mathbf{Sympathizes}_{i,j,H}\varphi \stackrel{def}{=}\ & \mathbf{IsSorryFor}_{i,j,H}\varphi\\ & \wedge \mathbf{Exp}_{i,j,H}\mathbf{AdoptGoal}_{i,j}\neg\varphi\end{aligned}$$

This definition logically entails the following theorem.

THEOREM 3.

$$\mathbf{Sympathizes}_{i,j,H}\varphi \rightarrow \mathbf{IsSaddened}_{i,j,H}\varphi$$

Thus, when agent i sympathizes with agent j about φ, it expresses that it is sad about φ.

4.2 Expression of complex emotions

In this section, we focus on expression of complex emotions (see Section 3.2). To express rejoicing is just to rejoice and to express gratitude is to thank (what corresponds to Vanderveken's definitions):

$$\mathbf{Rejoices}_{i,j,H}\varphi \stackrel{def}{=} \mathbf{Exp}_{i,j,H}\, Rejoicing_i\varphi$$

$$\mathbf{Thanks}_{i,j,H}\varphi \stackrel{def}{=} \mathbf{Exp}_{i,j,H}\, Gratitude_{i,j}\varphi$$

To rejoice and *to thank* both entail *to be delighted*.

THEOREM 4.

$$\mathbf{Rejoices}_{i,j,H}\varphi \rightarrow \mathbf{IsDelighted}_{i,j,H}\varphi \quad (4.1)$$

$$\mathbf{Thanks}_{i,j,H}\varphi \rightarrow \mathbf{IsDelighted}_{i,j,H}\varphi \quad (4.2)$$

To express regret is just to regret:

$$\mathbf{Regrets}_{i,j,H}\varphi \stackrel{def}{=} \mathbf{Exp}_{i,j,H}\mathit{Regret}_i\varphi$$

Following Vanderveken, *to deplore* is to express discontent with a high degree of strength and with a deep discontent or a deep sorrow. As we do not deal with degrees, *to deplore* is here just the expression of disappointment:

$$\mathbf{Deplores}_{i,j,H}\varphi \stackrel{def}{=} \mathbf{Exp}_{i,j,H}\mathit{Disappointment}_{i,j}\varphi$$

We can prove the following theorem.

THEOREM 5.

$$\mathbf{Regrets}_{i,j,H}\varphi \rightarrow \mathbf{IsSaddened}_{i,j,H}\varphi \quad (5.1)$$

$$\mathbf{Deplores}_{i,j,H}\varphi \rightarrow \mathbf{IsSaddened}_{i,j,H}\varphi \quad (5.2)$$

It means that if we regret for φ or if we deplore it, we are sad about the fact that φ is true.

Sometimes, we can also express some form of regret where the speaker is responsible for and where the consequence is bad for someone else. In this case, to express regret corresponds to *to apologize*. More precisely, agent i apologizes to agent j for φ if and only if, i expresses sadness about the fact that agent j did not achieve its goal that $\neg\varphi$ and i expresses that it believes to be responsible for φ:

$$\mathbf{Apologizes}_{i,j,H}\varphi \stackrel{def}{=} \mathbf{IsSaddened}_{i,j,H}(\mathbf{Goal}_j\neg\varphi \wedge \varphi) \wedge \mathbf{Exp}_{i,j,H}\mathbf{Bel}_i\,\mathbf{Resp}_i\varphi$$

This definition entails the following theorem.

THEOREM 6.

$$\mathbf{Apologizes}_{i,j,H}\varphi \rightarrow \mathbf{Regrets}_{i,j,H}(\mathbf{Goal}_j\neg\varphi \wedge \varphi)$$

Thus, when agent i apologizes to agent j for φ, i expresses regret about the fact that j has $\neg\varphi$ as a goal and φ is true.

The expression of moral satisfaction is defined as follows:

$$\mathbf{IsMorallySatisfied}_{i,j,H}\varphi \stackrel{def}{=} \mathbf{Exp}_{i,j,H}\mathit{MoralSatisfaction}_i\varphi$$

To express admiration is to compliment. Vanderveken says that "Complimenting does not necessarily relate to something done by the hearer, since we can compliment someone on his intelligence, musical ability (...)". But in these cases we can object that complimenting is more about the use of this intelligence or of this ability than about the intelligence itself or the ability itself. In any case, the following definition applies only to the case in which the hearer is responsible for φ:

$$\mathbf{Compliments}_{i,j,H}\varphi \stackrel{def}{=} \mathbf{Exp}_{i,j,H}\mathit{Admiration}_i\varphi$$

We can prove the following theorem.

THEOREM 7.

$$\mathbf{IsMorallySatisfied}_{i,j,H}\varphi \rightarrow \mathbf{ApprovesOf}_{i,j,H}\varphi \quad (7.1)$$

$$\mathbf{Compliments}_{i,j,H}\varphi \rightarrow \mathbf{ApprovesOf}_{i,j,H}\varphi \quad (7.2)$$

To express guilt is to express that one feels guilty, and to express reproach is just *to reproach*:

$$\mathbf{FeelsGuilty}_{i,j,H}\varphi \stackrel{def}{=} \mathbf{Exp}_{i,j,H}\mathit{Guilt}_i\varphi$$

$$\mathbf{Reproaches}_{i,j,H}\varphi \stackrel{def}{=} \mathbf{Exp}_{i,j,H}\mathit{Reproach}_{i,j}\varphi$$

These definitions entail the following theorem.

THEOREM 8.

$$\mathbf{FeelsGuilty}_{i,j,H}\varphi \rightarrow \mathbf{DisapprovesOf}_{i,j,H}\varphi \quad (8.1)$$

$$\mathbf{Reproaches}_{i,j,H}\varphi \rightarrow \mathbf{DisapprovesOf}_{i,j,H}\varphi \quad (8.2)$$

In other words, if agent i expresses that it feels guilty about the fact that φ is true, or if agent i reproaches agent j for φ, then agent i also expresses its disapproval for φ.

To accuse is not an expressive (but an assertive —see[26, p. 179]). It is however interesting to give a name to the expression of a speaker's belief about the hearer's responsibility:[13]

$$\mathbf{Accuses}_{i,j,H}\varphi \stackrel{def}{=} \mathbf{Exp}_{i,j,H}\mathbf{Bel}_i\,\mathbf{Resp}_j\varphi$$

We are now able to formalize the expressive *to protest*. Following Vanderveken, *to protest* is nothing but to express his/her disapproval together with the fact that the addressee of the act is responsible for the present state of affairs. The latter is what we call *to accuse*. Thus:

$$\mathbf{Protests}_{i,j,H}\varphi \stackrel{def}{=} \mathbf{DisapprovesOf}_{i,j,H}\varphi \wedge \mathbf{Accuses}_{i,j,H}\varphi$$

4.3 Remark

When the performance of an expressive entails the performance of another expressive – this is typically the case in the previous theorems –, it means that each time we express some psychological attitude, we also express some other psychological attitude. This relation exists in speech act theory through the semantic tree of expressives (see [26, p. 218]). In this tree, the success conditions of *to express* are a subset of the success conditions of *to approve*, and the success conditions of *to approve* are themselves a subset of success conditions of *to praise*, for instance. This means that, from an illocutionary point of view, *to praise* entails *to approve*, and *to approve* entails *to express*.

If we suppose that the speaker has the psychological attitudes that he/her expresses, then the previous theorems suggest that feeling some emotions entails feeling some others. For example, Theorem 5.1 says that feeling regret entails feeling sadness. This is in accordance with the literature in psychology according to which we can feel several emotions at the same time (see [13] for more details).

5. CONCLUSION

In this article we have presented the logic **MLC** that allows us to represent the cognitive structure of basic emotions (such as joy or sadness) and more complex emotions (such as regret or guilt), and their expression in front of a group of

[13]According to Vanderveken, when agent i accuses agent j of the fact that φ is true, agent i presupposes that φ is bad. This property needs the introduction of a new operator, but we do not intend here to give a subtle definition of this assertive: we just intend here to give a name to a particular formula of the language.

agents. Recall that a cognitive structure of emotion corresponds to the mental states that an agent must necessarily have for feeling the corresponding emotion.

Our work is based on the assumption that the performance of an illocutionary act consists in the expression of some mental states by the speaker. The logic **MLC** includes a novel modal operator formalizing what is expressed by performing a speech act. This operator allows us to formalize every class of illocutionary act. In this work, we only presented expressive speech acts because this class is less studied than the others (assertives, directives, commissives and declaratives). In future work, we will present a generalization of this work by including other classes of illocutionary acts.

By means of the logic **MLC** we have proved some intuitive theorems highlighting the relationships between different emotions (*e.g.* regret entails sadness) and between different expressive speech acts (*e.g.* to apologize entails to regret).

Note that we did not exploit in detail the argument H (the audience) in our formalization of expressive speech acts. However, as we have briefly shown in Section 4, the argument H becomes useful when we want to describe a private conversation within a group discussion. For instance, if a lecturer tells to the chairman that he/she has stage fright, there is no reason to suppose that every person who is present at the conference hears that. The argument H in the modal operator $\mathbf{Exp}_{i,j,H}$ allows us to represent such cases.

6. ACKNOWLEDGMENTS

This work has been supported by the French ANR project CECIL "Complex Emotions in Communication, Interaction and Language", contract No. ANR-08-CORD-005.

7. REFERENCES

[1] C. Adam, A. Herzig, and D. Longin. A logical formalization of the OCC theory of emotions. *Synthese*, 168:201–248, 2009.

[2] J. L. Austin. *How To Do Things With Words*. Oxford University Press, 1962.

[3] P. Blackburn, M. de Rijke, and Y. Venema. *Modal Logic*. Cambridge University Press, Cambridge, 2001.

[4] H. N. Castaneda. *Thinking and Doing*. D. Reidel, Dordrecht, 1975.

[5] H. Chockler and J. Y. Halpern. Responsibility and blame: a structural-model approach. *Journal of Artificial Intelligence Research*, 22(1):93–115, 2004.

[6] P. Cohen, J. Morgan, and M. Pollack, editors. *Intentions in communication*. The MIT Press, 1990.

[7] P. R. Cohen and H. J. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42:213–261, 1990.

[8] R. Conte and C. Castelfranchi. *Cognitive and social action*. London University College of London Press, London, 1995.

[9] F. Dignum and H. Weigand. Communication and deontic logic. In R. Wieringa and R. Feenstra, editors, *Information Systems, Correctness and Reusability*, pages 242–260. World Scientific, 1995.

[10] A. Herzig and D. Longin. A logic of intention with cooperation principles and with assertive speech acts as communication primitives. In *Proceedings of AAMAS 2002*, pages 920–927. ACM Press, 2002.

[11] J. F. Horty. *Agency and Deontic Logic*. Oxford University Press, Oxford, 2001.

[12] D. Kahneman and D. T. Miller. Norm theory: comparing reality to its alternatives. *Psychological Review*, 93:136–153, 1986.

[13] R. S. Lazarus. *Emotion and adaptation*. Oxford University Press, New York, 1991.

[14] E. Lorini and A. Herzig. A logic of intention and attempt. *Synthese*, 163(1):45–77, 2008.

[15] E. Lorini and F. Schwarzentruber. A logic for reasoning about counterfactual emotions. *Artificial Intelligence*, 175(3-4):814–847, 2011.

[16] A. Ortony, G. Clore, and A. Collins. *The cognitive structure of emotions*. Cambridge University Press, Cambridge, MA, 1988.

[17] A. S. Rao and M. P. Georgeff. Modelling rational agents within a BDI-architecture. In *Proceedings of KR'91*, pages 473–484. Morgan Kaufmann Publishers, 1991.

[18] J. A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178, 1980.

[19] K. R. Scherer and P. Ekman. *Approaches to emotion*. Erlbaum, Hillsdale, NJ, 1984.

[20] J. R. Searle. *Speech acts: An essay in the philosophy of language*. Cambridge University Press, New York, 1969.

[21] J. R. Searle. *Expression and Meaning. Studies on the Theory of Speech Acts*. Cambridge University Press, 1979.

[22] J. R. Searle. *Rationality in Action*. MIT Press, Cambridge, 2001.

[23] M. P. Singh. An ontology for commitments in multiagent systems. *Artificial Intelligence and Law*, 7:97–113, 1999.

[24] B. R. Steunebrink, M. Dastani, and J.-J. C. Meyer. The OCC model revisited. In D. Reichardt, editor, *Proceedings of the 4th Workshop on Emotion and Computing*, 2009.

[25] P. Turrini, J.-J. C. Meyer, and C. Castelfranchi. Coping with shame and sense of guilt: a dynamic logic account. *Journal of Autonomous Agents and Multi-Agent Systems*, 20(3), 2010.

[26] D. Vanderveken. *Principles of language use*, volume 1 of *Meaning and Speech Acts*. Cambridge University Press, 1990.

[27] M. Verdicchio and M. Colombetti. A logical model of social commitment for agent communication. In *Proceedings of AAMAS 2003*, pages 528–535. ACM Press, 2003.

[28] D. N. Walton and E. C. Krabbe. *Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning*. State University of New-York Press, NY, 1995.

[29] M. Zeelenberg, W. W. Van Dijk, and A. S. R. Manstead. Reconsidering the relation between regret and responsibility. *Organizational Behavior and Human Decision Processes*, 74:254–272, 1998.

A Logical Framework for Trust-Related Emotions

Jean-François Bonnefon[1], Dominique Longin[2] and Manh-Hung Nguyen[3]

[1] bonnefon@univ-tlse2.fr
CLLE, Université de Toulouse, CNRS, France

[2] Dominique.Longin@irit.fr
IRIT, Université de Toulouse, CNRS, France

[3] Manh-Hung.Nguyen@irit.fr
IRIT, Université de Toulouse, UPS, France

Abstract

Emotion and trust are two important concerns for the elaboration of interaction systems that would be closer and more attractive to their users, in particular by endowing machines with the ability to predict, understand, and process emotions and trust. This paper attempts to construct a common logical framework for the representation of emotion and trust. This logical framework combines a logic of belief and choice, a logic of time, and a dynamic logic. Using this common framework, we identify formal relations between trust and emotions, for which we also provide behavioral validation.

Keywords: Modal logic, emotions, trust, distrust.

1 Introduction

The rapidly growing field of affective computing aims at developing interaction systems that are closer and more attractive to their users, in particular by endowing machines with the ability to predict, understand, and process emotions (on the one hand), and trust (on the other hand). In this article, we introduce a unified logical approach to represent the cognitive structure of some emotions, of trust/distrust, and their relations at a formal level.

We formalize the concepts of emotions as well as trust/distrust based on cognitive models proposed by cognitive psychologists. Regarding emotions, we draw on cognitive theories (for more detail, see [26]) which assume that emotions are closely tied to changes in beliefs and desires. We capitalize on psychological models that allow to recognize and distinguish emotions based on their decomposition in cognitive factors particularly the cognitive structure of emotion of Ortony et al. [22], the cognitive patterns of emotion of Lazarus [20] and the belief-desire theory

of emotion (BDTE) [24, 10]. Similarly, we attempt to adhere closely to cognitive definition of trust [4] and distrust [5].

Although there are tight conceptual connections between emotion and trust [19], and although there ware some separated formalization of the concepts of trust as the works of Herzig et al. [17], and the concepts of emotions such as the works of Adam et al. [1] and Steunebrink et al. [28, 27], there is not yet a common logic to represent them both. Our work aims at filling that gap by formally representing trust and emotions in a common logic; this common logic will enable us to lay bare the formal relations between trust and emotion. The logic we offer is a combination of the logic of beliefs and choices as the one of Herzig and Longin [16] (a refinement from Cohen and Levesque [8]), the logic of time (introduced by Arthur Prior [23]), and dynamic logic introduced by Fischer and Ladner [11] and Harel et al. [15].

This paper is organized as follows: Part 2 introduces the logical framework. Part 3 formalizes the cognitive structure of some emotions, Part 4 formalizes the cognitive structure of trust and distrust. Part 5 shows some formal relations in the effect of trust/distrust on the emotions, and provides behavioral validation for these relations.

2 Logical Framework

Syntax. The syntactic primitives of our logic are as follows: a nonempty finite set of agents $AGT = \{i_1, i_2, \ldots, i_n\}$, a nonempty finite set of atomic events $EVT = \{e_1, e_2, \ldots, e_p\}$, and a nonempty set of atomic propositions $ATM = \{p_1, p_2, \ldots\}$. The variables i, j, k... denote agents. The expression $i_1{:}e_1 \in AGT \times EVT$ denotes an event e_1 intentionally caused by agent i_1 and e_1 is thus called an "action". The variables α, β... denote such actions. The language of our logic is defined by the following BNF :

$$\varphi ::= p \mid i{:}\alpha\text{-}happens \mid \neg\varphi \mid \varphi \vee \varphi \mid \mathtt{X}\varphi \mid \mathtt{X}^{-1}\varphi \mid \mathtt{G}\varphi \mid \mathtt{Bel}_i\,\varphi \mid \mathtt{Choice}_i\,\varphi \mid \mathtt{Grd}_I\varphi$$

where p ranges over ATM, $i{:}\alpha$ ranges over $AGT \times EVT$, $i{:}\alpha\text{-}happens$ ranges over ATM for each $i{:}\alpha \in AGT \times EVT$, and $I \subseteq AGT$. The classical boolean connectives $\wedge$ (conjunction), $\rightarrow$ (material implication), $\leftrightarrow$ (material equivalence), $\top$ (tautology) and $\bot$ (contradiction) are defined from $\neg$ (negation) and $\vee$ (disjunction).

$i{:}\alpha\text{-}happens$ reads "agent i is just about to perform the action α"; $\mathtt{X}\varphi$ reads "φ will be true next instant"; $\mathtt{X}^{-1}\varphi$ reads "φ was true at the previous instant"; $\mathtt{G}\varphi$ reads "henceforth, φ is true"; $\mathtt{Bel}_i\,\varphi$ reads "agent i believes that φ is true"; $\mathtt{Choice}_i\,\varphi$ reads "agent i prefers that φ be true"; $\mathtt{Grd}_I\varphi$ reads "φ is publicly grounded between the agents in group I".

Note 1 *In the workshop, we used the concept of common belief (or mutual belief)* $\mathtt{MBel}_{i,j}$. *This concept is usually defined as follows:*

$$\mathtt{MBel}_{i,j}\phi \stackrel{def}{=} \mathtt{Bel}_i\,\phi \wedge \mathtt{Bel}_j\,\phi \wedge \mathtt{Bel}_i\,\mathtt{Bel}_j\,\phi \wedge \mathtt{Bel}_j\,\mathtt{Bel}_i\,\phi \ldots$$

Thus, in KD45 logic for Bel, $\mathtt{Bel}_i\,\mathtt{MBel}_{i,j}\phi \rightarrow \mathtt{MBel}_{i,j}\phi$ *(and the same for* $\mathtt{Bel}_j$ *).*

This is problematic for trust, because the fact that $\mathtt{Bel}_i\,\mathtt{MBel}_{i,j}\phi$ (i.e., *the fact that the truster* ***believes*** *that there is a common belief about* ϕ *between* i *and* j*)*

be true is a sufficient condition. $\texttt{MBel}_{i,j}\phi$ *is thus too strong because it necessarily implies that agent* i *cannot be wrong (when* i *believes that there is a common belief about* ϕ*). Maybe* i *is wrong, but it is sufficient.*

Thus, we need an operator $\Box$ *whose the meaning is "*ϕ *is okay for* i *and* j*" such that* $\texttt{Bel}_i\,\Box\phi \not\rightarrow \Box\phi$*. This operator is the new operator of grounding we have introduced in the final version of our article.*

This operator $\Box$ *is the new operator* $\texttt{Grd}_I$ *where* $I \subseteq AGT$*. This operator is very close to the concept of group belief used in Walton & Krabbe (1995).*

We define the following abbreviations:

$$i{:}\alpha\text{-}done \stackrel{def}{=} \texttt{X}^{-1} i{:}\alpha\text{-}happens \qquad (\text{Def}_{i:\alpha\text{-}done})$$

$$\texttt{Happens}_{i:\alpha}\varphi \stackrel{def}{=} i{:}\alpha\text{-}happens \wedge \texttt{X}\varphi \qquad (\text{Def}_{\texttt{Happens}_{i:\alpha}})$$

$$\texttt{After}_{i:\alpha}\varphi \stackrel{def}{=} i{:}\alpha\text{-}happens \rightarrow \texttt{X}\varphi \qquad (\text{Def}_{\texttt{After}_{i:\alpha}})$$

$$\texttt{Done}_{i:\alpha}\varphi \stackrel{def}{=} i{:}\alpha\text{-}done \wedge \texttt{X}^{-1}\varphi \qquad (\text{Def}_{\texttt{Done}_{i:\alpha}})$$

$$\texttt{F}\varphi \stackrel{def}{=} \neg\texttt{G}\neg\varphi \qquad (\text{Def}_{\texttt{F}})$$

$$\texttt{Goal}_i\,\varphi \stackrel{def}{=} \texttt{Choice}_i\,\texttt{FBel}_i\,\varphi \qquad (\text{Def}_{\texttt{Goal}_i})$$

$$\texttt{Intend}_i\,\alpha \stackrel{def}{=} \texttt{Choice}_i\,\texttt{F}i{:}\alpha\text{-}happens \qquad (\text{Def}_{\texttt{Intend}_i})$$

$$\texttt{Capable}_i\alpha \stackrel{def}{=} \neg\texttt{After}_{i:\alpha}\bot \qquad (\text{Def}_{\texttt{Capable}_i})$$

$$\texttt{Possible}_i\varphi \stackrel{def}{=} \neg\texttt{Bel}_i\,\neg\varphi \qquad (\text{Def}_{\texttt{Possible}_i})$$

$$\texttt{Awareness}_i\varphi \stackrel{def}{=} \texttt{X}^{-1}\neg\texttt{Bel}_i\,\varphi \wedge \texttt{Bel}_i\,\varphi \qquad (\text{Def}_{\texttt{Awareness}_i})$$

$i{:}\alpha\text{-}done$ reads "agent i has done action α"; $\texttt{Happens}_{i:\alpha}\varphi$ reads "agent i is doing action α and φ will be true next instant"; $\texttt{After}_{i:\alpha}\varphi$ reads "φ is true after any execution of α by i"; $\texttt{Done}_{i:\alpha}\varphi$ reads "agent i has done action α and φ was true at previous instant"; $\texttt{F}\varphi$ reads "φ will be true in some future instants"; $\texttt{Goal}_i\,\varphi$ reads "agent i has the goal (chosen preference) that φ be true"; $\texttt{Intend}_i\,\alpha$ reads "agent i intends to do α"; $\texttt{Capable}_i\alpha$ reads "agent i is capable to do α"; $\texttt{Possible}_i\varphi$ reads "agent i believes that it is possible φ"; $\texttt{Awareness}_i\varphi$ reads "agent i has just experienced that φ is true".

Semantics. For temporal operators, we use a semantics based on linear time described by a sequence (or story) of time points. (This semantics is very close to CTL* [7]) A frame $\mathcal{F}$ is a 4-tuples $\langle H, \mathscr{B}, \mathscr{C}, \mathscr{G}\rangle$ where: H is a set of stories that are represented as sequences of time points, where each time point is identified by an integer $z \in \mathbb{Z}$, a time point z in a story h is called a situation $< h, z >$; $\mathscr{B}$ is the set of all $\mathscr{B}_i$ such that $\mathscr{B}_i(h, z)$ denotes the set of stories believed as being possible by the agent i in the situation $< h, z >$; $\mathscr{C}$ is the set of all $\mathscr{C}_i$ such that $\mathscr{C}_i(h, z)$ denotes the set of stories chosen by the agent i in the situation $< h, z >$; $\mathscr{G}$ is the set of all $\mathscr{G}_I$ such that $\mathscr{G}_I(h, z)$ denotes the set of stories which are publicly grounded in the group I of agents, in the situation $< h, z >$.

All the accessibility relations $\mathscr{B}_i$ are serial[1], transitive[2] and euclidean[3]. This semantic is completely standard in epistemic logic (see [18, 14]) All the accessibility relations $\mathscr{G}_I$ are serial, transitive and euclidean (This is similar to the operator group grounding introduced by Gaudou et al. [13]). All the accessibility $\mathscr{C}_i$ are serial. Moreover, we impose for every $z \in \mathbb{Z}$ that: if $h' \in \mathscr{B}_i(h, z)$ then $\mathscr{C}_i(h, z) = \mathscr{C}_i(h', z)$. It means that if an agent believes that the world h' is possible from the world h, then the set of his/her preference worlds from h and h' are the same. In other terms, the worlds an agent prefers and the ones that agent believes that s/he prefers are the same (briefly, the agent is conscious about his/her preferences, and s/he prefers what s/he believes that s/he prefers).

A model $\mathcal{M}$ is a couple $\langle \mathcal{F}, \mathcal{V} \rangle$ where $\mathcal{F}$ is a frame and $\mathcal{V}$ is a function associating each atomic proposition p with the set $\mathcal{V}(p)$ of couple (h, z) where p is true. Truth conditions are defined as follows:

$\mathcal{M}, h, z \models p$ iff $(h, z) \in \mathcal{V}(p)$
$\mathcal{M}, h, z \models \mathtt{X}\varphi$ iff $\mathcal{M}, h, z + 1 \models \varphi$
$\mathcal{M}, h, z \models \mathtt{X}^{-1}\varphi$ iff $\mathcal{M}, h, z - 1 \models \varphi$
$\mathcal{M}, h, z \models \mathtt{G}\varphi$ iff $\mathcal{M}, h, z' \models \varphi$ for every $z' \geq z$
$\mathcal{M}, h, z \models \mathtt{Bel}_i\, \varphi$ iff $\mathcal{M}, h', z \models \varphi$ for every $(h', z) \in \mathscr{B}_i(h, z)$
$\mathcal{M}, h, z \models \mathtt{Choice}_i\, \varphi$ iff $\mathcal{M}, h', z \models \varphi$ for every $(h', z) \in \mathscr{C}_i(h, z)$
$\mathcal{M}, h, z \models \mathtt{Grd}_I\varphi$ iff $\mathcal{M}, h', z \models \varphi$ for every $(h', z) \in \mathscr{G}_I(h, z)$

Other truth conditions are defined as usual.

Axiomatics. Due to our linear time semantics, the temporal operators satisfy the following principles:

$$i{:}\alpha\text{-}happens \leftrightarrow \mathtt{X}i{:}\alpha\text{-}done \tag{1}$$
$$\mathtt{X}\varphi \leftrightarrow \neg\mathtt{X}\neg\varphi \tag{2}$$
$$\varphi \leftrightarrow \mathtt{X}\mathtt{X}^{-1}\varphi \tag{3}$$
$$\varphi \leftrightarrow \mathtt{X}^{-1}\mathtt{X}\varphi \tag{4}$$
$$\mathtt{G}\varphi \leftrightarrow \varphi \wedge \mathtt{X}\mathtt{G}\varphi \tag{5}$$
$$\mathtt{G}(\varphi \rightarrow \mathtt{X}\varphi) \rightarrow (\varphi \rightarrow \mathtt{G}\varphi) \tag{6}$$

$\mathtt{Bel}_i$ and $\mathtt{Choice}_i$ operators are defined in a normal modal logic plus (D) axioms.[4] Thus, if $\Box$ represents a $\mathtt{Bel}_i$ operator or $\mathtt{Choice}_i$ operator:

$$\frac{\varphi}{\Box\varphi} \tag{RN$_\Box$}$$
$$\Box(\varphi \rightarrow \psi) \rightarrow (\Box\varphi \rightarrow \Box\psi) \tag{K$_\Box$}$$
$$\Box\varphi \rightarrow \neg\Box\neg\varphi \tag{D$_\Box$}$$

For example, axiom D$_\Box$ applied to operator $\mathtt{Bel}_i$ is $D_{\mathtt{Bel}_i}$, which is described as: $\mathtt{Bel}_i\, \varphi \rightarrow \neg\mathtt{Bel}_i\, \neg\varphi$.

[1] for every $i \in AGT$, $\mathscr{B}_i(h, z) \neq \emptyset$
[2] if $< h', z > \in \mathscr{B}_i(h, z)$ and $< h'', z > \in \mathscr{B}_i(h', z)$, then $< h'', z > \in \mathscr{B}_i(h, z)$
[3] if $< h', z > \in \mathscr{B}_i(h, z)$ and $< h'', z > \in \mathscr{B}_i(h, z)$, then $< h'', z > \in \mathscr{B}_i(h', z)$
[4] We use here the notation of modal logic axioms introduced by Chellas in [6].

($\mathrm{RN}_{\Box}$) means that all theorems are believed (respectively: chosen) by every agent i; ($\mathrm{K}_{\Box}$) means that beliefs (respectively: choices) are closed under material implication for every agent i; ($\mathrm{D}_{\Box}$) means that beliefs (respectively: choices) of every agent i are rational: they cannot be contradictory.

The $\mathtt{Bel}_i$ operators satisfy the following principles of introspection:

$$\mathtt{Bel}_i\,\varphi \leftrightarrow \mathtt{Bel}_i\,\mathtt{Bel}_i\,\varphi \qquad (4_{\mathtt{Bel}_i})$$

$$\neg\mathtt{Bel}_i\,\varphi \leftrightarrow \mathtt{Bel}_i\,\neg\mathtt{Bel}_i\,\varphi \qquad (5_{\mathtt{Bel}_i})$$

that means that agent i is conscious of its beliefs and of its disbeliefs.

The following principle follows from the semantical constraint between belief accessibility relation and choice accessibility relation, and from axiom ($\mathrm{D}_{\Box}$) for $\mathtt{Bel}_i$:

$$\mathtt{Choice}_i\,\varphi \leftrightarrow \mathtt{Bel}_i\,\mathtt{Choice}_i\,\varphi \qquad (4_{BC})$$

$$\neg\mathtt{Choice}_i\,\varphi \leftrightarrow \mathtt{Bel}_i\,\neg\mathtt{Choice}_i\,\varphi \qquad (5_{BC})$$

that means that agent i is conscious of its choices and of its dischoices.

The sound and complete axiomatization of $\mathtt{Grd}_I$ operator is defined as the one of common belief operator (also called mutual belief), which is closed to the operator described in Walton and Krabbe [29], also introduced by Gaudou et al. [13]:

$$\frac{\varphi}{\mathtt{Grd}_I\varphi} \qquad (\mathrm{RN}_{\mathtt{Grd}_I})$$

$$\mathtt{Grd}_I(\varphi \rightarrow \psi) \rightarrow (\mathtt{Grd}_I\varphi \rightarrow \mathtt{Grd}_I\psi) \qquad (\mathrm{K}_{\mathtt{Grd}_I})$$

$$\mathtt{Grd}_I\varphi \rightarrow \neg\mathtt{Grd}_I\neg\varphi \qquad (\mathrm{D}_{\mathtt{Grd}_I})$$

$$\mathtt{Grd}_I\varphi \rightarrow \mathtt{Grd}_I\mathtt{Grd}_I\varphi \qquad (4_{\mathtt{Grd}_I})$$

$$\neg\mathtt{Grd}_I\varphi \rightarrow \mathtt{Grd}_I\neg\mathtt{Grd}_I\varphi \qquad (5_{\mathtt{Grd}_I})$$

Axiom ($\mathrm{RN}_{\mathtt{Grd}_I}$) means that every tautology is public ground. Axiom ($\mathrm{K}_{\mathtt{Grd}_I}$) means that if φ is publicly grounded in I and that φ implies ψ then ψ is also publicly grounded in I. Axiom ($\mathrm{D}_{\mathtt{Grd}_I}$) means that the set of grounded informations is consistent: it can not be the case that both φ and $\neg\varphi$ are simultaneously grounded. The positive introspection axiom ($4_{\mathtt{Grd}_I}$) and negative introspection axiom ($5_{\mathtt{Grd}_I}$) account for the public character of $\mathtt{Grd}_I$. From these collective awareness results: if φ has (resp. has not) been grounded then it is established that φ has (resp. has not) been grounded.

Linear time semantics entail the following principles:

$$\mathtt{G}\varphi \rightarrow \mathtt{After}_{i:\alpha}\varphi \qquad (7)$$

$$\mathtt{Happens}_{i:\alpha}\varphi \rightarrow \mathtt{After}_{j:\beta}\varphi \qquad (8)$$

$$\mathtt{After}_{i:\alpha}\varphi \leftrightarrow \neg\mathtt{Happens}_{i:\alpha}\neg\varphi \qquad (9)$$

Axiom (7) describe the relationship between time and action: if henceforth φ is true then after every action α of every agent i, φ will be true. (Note that the converse is not valid: it is possible that φ be true after every action α of every agent i performed in a situation $< h, z >$, and that φ be false at time $z' > z$.)

As time is linear, actions are deterministic on a given history. Thus, axiom (8) reads: if agent i is just about to perform α after what φ will be true, then after every performance of every action β by every agent j, φ will be true. In other words,

if action α leads to a time point where φ is true, then every action performed by every agent leads to this time point.

Finally, axiom (9) means that $\texttt{After}_{i:\alpha}$ and $\texttt{Happens}_{i:\alpha}$ operators are dual operators. This property is fair with respect to dynamic logic [15].

3 Formalization of the cognitive structure of emotion

In this section, we present the formalization of emotions, based on their cognitive structure as proposed by Ortony et al. [22], Frijda [12] as well as those of Reisenzei [24] and Scherer et al. [25].

Joy/Distress. The cognitive structure of *Joy* consists of two main factors: (i) a proposition φ is desirable for agent i, and (ii) agent i just experienced that φ is the case. To formalize the first factor, we consider that agent i desiring φ means that i wants φ to be the case. So we formalize desire as a goal (chosen preference). Therefore, the first factor is potentially formalized as $\texttt{Goal}_i\,\varphi$, the second factor may be formalized as $\texttt{Bel}_i\,\varphi$.

However, we assume that emotion is triggered at the moment when all its factors are fulfilled, and that its intensity then decreases with time [9, 12]. Accordingly, we include a time factor into most emotional formulas. Thus, the first factor of *Joy* in particular means that agent i now recalls that at the previous instant, s/he desired φ, until experiencing that φ was in fact true: $\texttt{Bel}_i\,\texttt{X}^{-1}\texttt{Goal}_i\,\varphi$. It means that in order to be joyful, agent i must keep in mind his desire in the previous instant. Hereafter, we add this analysis for almost emotional formulas. The second factor means that agent i has just experienced that φ is true and did not previously know it: $\texttt{Awareness}_i\varphi$.

The same analysis applies to *Distress*, except that in the first factor of *Distress*, φ is undesirable for agent i, which we assume to mean that agent i desired $\neg\varphi$: $\texttt{Bel}_i\,\texttt{X}^{-1}\texttt{Goal}_i\,\neg\varphi$. We accordingly formalize the concept of *Joy* and *Distress*:

Definition 1 (Joy/Distress)

$$\texttt{Joy}_i\,\varphi \stackrel{def}{=} \texttt{Bel}_i\,\texttt{X}^{-1}\texttt{Goal}_i\,\varphi \wedge \texttt{Awareness}_i\varphi$$

$$\texttt{Distress}_i\,\varphi \stackrel{def}{=} \texttt{Bel}_i\,\texttt{X}^{-1}\texttt{Goal}_i\,\neg\varphi \wedge \texttt{Awareness}_i\varphi$$

To illustrate the definition of *Joy*, we can say that an individual is joyful when he has just realized that he won the lottery ($\texttt{Awareness}_{man}$*(win lottery)*) with the trivial assumption that he had been desiring to win the lottery ($\texttt{Bel}_{man}\,\texttt{X}^{-1}\,\texttt{Goal}_{man}$ *(win lottery)*). In contrast, to illustrate the definition of *Distress*, we can say that an individual feels distress when she learns she has lost her job ($\texttt{Awareness}_{woman}$*(lost job)*) assuming that she had the goal not to lose her job ($\texttt{Bel}_{woman}\,\texttt{X}^{-1}\texttt{Goal}_{woman}\,\neg$*(lost job)*).

Hope/Fear. The cognitive structure of *Hope* consists of two factors: (i) a proposition φ is desirable for agent i, and (ii) agent i believes that φ may be true in the future. To formalize the first factor, we consider that φ is not true at the moment when i hopes for it: $\texttt{Goal}_i\,\varphi$.

We interpret the second factor, as meaning that among all of possible future worlds, agent i believes that there is at least one world in which φ will be the case. In other terms, agent i does not believe that φ will be false in all of possible future worlds: $\texttt{Possible}_i\texttt{F}\varphi$. If i believes that φ can never be the case in all of possible future worlds, then i has no ground for hope.

The same analysis applies to *Fear*, except that φ is now undesirable for agent i: $\texttt{Goal}_i\ \neg\varphi$.

We accordingly formalize the concept of *Hope* and *Fear* as follows:

Definition 2 (Hope/Fear)

$$\texttt{Hope}_i\ \varphi \stackrel{def}{=} \texttt{Goal}_i\ \varphi \wedge \texttt{Possible}_i\texttt{F}\varphi$$
$$\texttt{Fear}_i\ \varphi \stackrel{def}{=} \texttt{Goal}_i\ \neg\varphi \wedge \texttt{Possible}_i\texttt{F}\varphi$$

For example, a debutante is hopeful about being asked to dance, for she thinks it is possible ($\texttt{Possible}_{girl}\texttt{F}$*(being asked to dance)*) and this is what she wants ($\texttt{Goal}_{girl}$ *(being asked to dance)*). In contrast, an employee fears to be fired when he does not wish to be fired ($\texttt{Goal}_{employee}\ \neg(fired)$) but believes it is a possibility $\texttt{Possible}_{employee}\texttt{F}$*(to be fired)*).

Satisfaction/Disappointment. The cognitive structure of *Satisfaction* consists of three factors: (i) agent i desired a proposition φ, (ii) agent i used to believe that φ might be true in the near future, and (iii) agent i now experiences that φ is really the case. The first two factors mean that now, agent keeps in mind that at the previous instant, s/he desired φ and believed that φ could be true in the future ($\texttt{Bel}_i\ \texttt{X}^{-1}(\texttt{Goal}_i\ \varphi \wedge \texttt{Possible}_i\texttt{F}\varphi)$) (cf. the analysis of the second factor of *Hope*). The last factor means that i now experiences that φ is true, but did not know it the previous instant ($\texttt{Awareness}_i\varphi$).

The difference in the case of *Disappointment* is agent recalls that, in the previous instant, s/he desired $\neg\varphi$ instead of φ, and s/he believed that $\neg\varphi$ was possibly true in the future ($\texttt{Bel}_i\ \texttt{X}^{-1}(\texttt{Goal}_i\ \neg\varphi \wedge \texttt{Possible}_i\texttt{F}\neg\varphi)$). We formalize *Satisfaction* and *Disappointment* as follows:

Definition 3 (Satisfaction/Disappointment)

$$\texttt{Satisfaction}_i\ \varphi \stackrel{def}{=} \texttt{Bel}_i\ \texttt{X}^{-1}(\texttt{Goal}_i\ \varphi \wedge \texttt{Possible}_i\texttt{F}\varphi) \wedge \texttt{Awareness}_i\varphi$$
$$\texttt{Disappointment}_i\ \varphi \stackrel{def}{=} \texttt{Bel}_i\ \texttt{X}^{-1}(\texttt{Goal}_i\ \neg\varphi \wedge \texttt{Possible}_i\texttt{F}\neg\varphi) \wedge \texttt{Awareness}_i\varphi$$

For example, when the debutante realizes that she is indeed asked to dance ($\texttt{Awareness}_{girl}$*(asked to dance)*) she is satisfied. Were she not to be asked to dance
($\texttt{Awareness}_{girl}$*(not asked to dance)*), she would feel disappointed.

We can point out the relations between *Satisfaction*, *Disappointment* and *Hope*:

$$\texttt{Satisfaction}_i\ \varphi \leftrightarrow \texttt{Bel}_i\ \texttt{X}^{-1}\texttt{Hope}_i\ \varphi \wedge \texttt{Awareness}_i\varphi \quad (10)$$
$$\texttt{Disappointment}_i\ \varphi \leftrightarrow \texttt{Bel}_i\ \texttt{X}^{-1}\texttt{Hope}_i\ \neg\varphi \wedge \texttt{Awareness}_i\varphi \quad (11)$$

The relation between *Satisfaction* and *Joy* can be formalized as the following proposition:

Proposition 1 (Satisfaction implies Joy)

$$\texttt{Satisfaction}_i\,\varphi \rightarrow \texttt{Joy}_i\,\varphi$$

That is, if we feel satisfaction about something, then we will also feel joy about it.

Fear-confirmed/Relief. The cognitive structure of *Fear-confirmed* consists of three factors: (i) a proposition φ was undesirable for agent i, (ii) agent i believed that φ might be true in the near future, and (iii) agent i now experiences that φ is really true.

We use the same analysis as for *Satisfaction*, except agent recalls that in the previous instant, $\neg\varphi$ was desirable for agent i ($\texttt{Bel}_i\,\texttt{X}^{-1}\texttt{Goal}_i\,\neg\varphi$).

The difference in the case of *Relief* is agent recalls that, in the previous instant, s/he desired φ ($\texttt{Bel}_i\,\texttt{X}^{-1}\texttt{Goal}_i\,\varphi$), and believed that $\neg\varphi$ might be true in the near future ($\texttt{Bel}_i\,\texttt{X}^{-1}(\texttt{Goal}_i\,\varphi \wedge \texttt{Possible}_i\texttt{F}\neg\varphi)$). We formalize *Fear-confirmed* and *Relief* as:

Definition 4 (Fear-confirmed/Relief)

$$\texttt{FearConfirmed}_i\,\varphi \stackrel{def}{=} \texttt{Bel}_i\,\texttt{X}^{-1}(\texttt{Goal}_i\,\neg\varphi \wedge \texttt{Possible}_i\texttt{F}\varphi) \wedge \texttt{Awareness}_i\varphi$$

$$\texttt{Relief}_i\,\varphi \stackrel{def}{=} \texttt{Bel}_i\,\texttt{X}^{-1}(\texttt{Goal}_i\,\varphi \wedge \texttt{Possible}_i\texttt{F}\neg\varphi) \wedge \texttt{Awareness}_i\varphi$$

For example, the employee's fear of being fired is confirmed when he learns that he is indeed about to be fired ($\texttt{Awareness}_{employee}(fired)$) which he had been afraid of ($\texttt{Bel}_{employee}\,\texttt{X}^{-1}(\texttt{Goal}_{employee}\,\neg(fired) \wedge \texttt{Possible}_{employee}\texttt{F}(fired))$). In contrast, were he to learn that he is not going to be fired ($\texttt{Awareness}_{employee}$*(not fired)*), he would feel relief.

We can also point out the relations between *Fear-confirmed*, *Relief* and *Fear*:

$$\texttt{FearConfirmed}_i\,\varphi \leftrightarrow \texttt{Bel}_i\,\texttt{X}^{-1}\texttt{Fear}_i\,\varphi \wedge \texttt{Awareness}_i\varphi \tag{12}$$

$$\texttt{Relief}_i\,\varphi \leftrightarrow \texttt{Bel}_i\,\texttt{X}^{-1}\texttt{Fear}_i\,\neg\varphi \wedge \texttt{Awareness}_i\varphi \tag{13}$$

The relation between *Fear-confirmed* and *Distress* is stated in the following proposition:

Proposition 2 (Fear-confirmed implies Distress)

$$\texttt{FearConfirmed}_i\,\varphi \rightarrow \texttt{Distress}_i\,\varphi$$

That is, if our fears about something are confirmed, then we feel distressed.

4 Formalization of Trust

We now present the formalization of trust and distrust based on the cognitive definition of Castelfranchi and colleagues [4, 5].

Trust. We formalize the concept of trust based on Castelfranchi and Falcone's definition [4] of trust in action which says that agent i trusts agent j to ensure φ by performing action α if and only if agent i desires to achieve φ ($\texttt{Goal}_i\,\varphi$), and agent i expects that: (i) φ can be achieved by doing action α ($\texttt{Bel}_i\,\texttt{After}_{j:\alpha}\varphi$); (ii) agent j is able to perform action α ($\texttt{Bel}_i\,\texttt{Capable}_j\alpha$); and (iii) agent j has the intention to do such an action ($\texttt{Bel}_i\,\texttt{Intend}_j\,\alpha$).

However, these three factors are only necessary conditions, but not sufficient ones. For example, imagine that a robber wants to steal something located on the second floor of a mansion. There is a nurse on the first floor. The robber desires that the nurse stays where she is, because it makes his robbery possible. He also believes that it is possible that the nurse will stay where she is, and that it is actually her intention. Thus, the three conditions are satisfied, but we are reluctant nonetheless to say that the robber trusts the nurse to stay where she is in order to allow for his stealing, because there is no agreement between the nurse (trustee) and the robber (trustor). So here we need to add another condition for trust: an agreement between trustor and trustee that the trustee will perform such an action ($\texttt{Grd}_I trustee : \alpha\text{-}happens$), where $I = \{trustor, trustee\}$. We accordingly formalize the concept of trust as:

Definition 5 (Trust)

$$\texttt{Trust}_{i,j}(\alpha,\varphi) \stackrel{def}{=} \texttt{Goal}_i\,\varphi \wedge \texttt{Bel}_i\,\texttt{After}_{j:\alpha}\varphi \wedge \texttt{Bel}_i\,\texttt{Capable}_j\alpha \wedge \texttt{Bel}_i\,\texttt{Intend}_j\,\alpha \wedge \texttt{Grd}_{\{i,j\}}j{:}\alpha\text{-}happens$$

For example, a boss trusts his secretary to prepare a report in order to present it at a company meeting because the boss desires the report ($\texttt{Goal}_{boss}$ *(report)*), and in his opinion, the report can be possibly ready after the secretary prepares it ($\texttt{Bel}_{boss}\,After_{secretary:prepare}$*(report)*), the secretary has the ability and intention to prepare the report ($\texttt{Bel}_{boss}\,\texttt{Capable}_{secretary}(prepare) \wedge \texttt{Bel}_{boss}\,\texttt{Intend}_{secretary}\,(prepare)$)). It is clear that in the relation between the boss and his secretary, there is an agreement that the secretary will prepare the report in time ($\texttt{Grd}_{boss,secretary} secretary : prepare\text{-}happens$).

Distrust. We also adopt the definition of distrust given by Castelfranchi et al. [5] which says that agent i distrusts agent j to ensure φ by performing action α if and only if agent i desires to achieve φ ($\texttt{Goal}_i\,\varphi$), and agent i believes that at least one of these conditions is fulfilled: (i) agent j is not in the capacity to do action α: $\texttt{Bel}_i\,\neg\texttt{After}_{j:\alpha}\varphi$, or (ii) agent j is able to do α but he has not intention to do α: $\texttt{Possible}_i\texttt{After}_{j:\alpha}\varphi \wedge \texttt{Bel}_i\,\neg\texttt{Intend}_j\,\alpha$. We accordingly formalize this concept as:

Definition 6 (Distrust)

$$\texttt{DisTrust}_{i,j}(\alpha,\varphi) \stackrel{def}{=} \texttt{Goal}_i\,\varphi \wedge (\texttt{Bel}_i\,\neg\texttt{After}_{j:\alpha}\varphi \vee (\texttt{Possible}_i\texttt{After}_{j:\alpha}\varphi \wedge \texttt{Bel}_i\,\neg\texttt{Intend}_j\,\alpha))$$

For example, in spite of desiring the report ($\texttt{Goal}_{boss}$ *(report)*), the boss does not trust a new employee to prepare it because he believes the new employee is unable to perform that task($\texttt{Bel}_{boss}\,\neg\texttt{After}_{employee:prepare}$*(report)*).

From this definition, we can decompose the concept of distrust based only on the ability of trustee:

Definition 7 (Distrust based on ability)

$$\texttt{C-DisTrust}_{i,j}(\alpha,\varphi) \stackrel{def}{=} \texttt{Goal}_i\,\varphi \wedge \texttt{Bel}_i\,\neg\texttt{After}_{j:\alpha}\varphi$$

5 Trust-Related Emotions

5.1 Formal Relations

Trust and Hope. *Trust* and *Hope* have an important relation because they both feature a positive expectation [4]. When i trusts j, i has a positive expectation about j's power and performance. $Hope$ also implies some positive expectation. The greater the expectations, the deeper the trust; and, conversely, the deeper the disappointment when expectations are unrealized [3].

The first relation is formalized as follows:

Proposition 3 (Trust implies Hope)

$$\texttt{Trust}_{i,j}(\alpha,\varphi) \rightarrow \texttt{Hope}_i\varphi$$

This means that when we trust someone about an action that will bring some results, we are hopeful that the results will be obtained. For example, in a commercial transaction, when the buyer trusts his seller to send him a product after payment ($\texttt{Trust}_{buyer,seller}(send, receipt)$), he will be hopeful that he will receive the product ($\texttt{Hope}_{buyer}$ *receive product*). This proposition will be proved by applying Lemma 1: if we believe that φ is true after every execution of action α, and that someone is able to do α, then we believe that there is at least a future world in which φ is true.

Lemma 1

$$\texttt{Bel}_i\,\texttt{After}_{j:\alpha}\varphi \wedge \texttt{Bel}_i\,\texttt{Capable}_j\alpha \rightarrow \texttt{Possible}_i\texttt{F}\varphi$$

Once we trust someone to do an action to bring us something, we hope for the positive result of the action. In case of success, we feel satisfaction (formalized as Proposition 4). Conversely, in case of failure, we feel disappointment (formalized as Proposition 5).

Proposition 4 (Successful Trust implies Satisfaction)

$$\texttt{Bel}_i\,\texttt{Done}_{j:\alpha}\texttt{Trust}_{i,j}(\alpha,\varphi) \wedge \texttt{Awareness}_i\varphi \rightarrow \texttt{Satisfaction}_i\varphi$$

This means that when we believe that what we trusted has now occurred, we are satisfied about it. For example, when the boss trusted his secretary to prepare the report ($\texttt{Done}_{secretary:prepare}\texttt{Trust}_{boss,secretary}(prepare,$ *having report*$)$), and on the morning of the day after, he has received the report ($\texttt{Bel}_{boss}$ *having report*), then he is satisfied ($\texttt{Satisfaction}_{boss}$*having report*). This proposition has a corollary which is deduced from Proposition 1 and 4: When we experience that what we trusted has really occurred, we will also feel joy about it.

Corollary 1

$$\texttt{Bel}_i\,\texttt{Done}_{j:\alpha}\texttt{Trust}_{i,j}(\alpha,\varphi) \wedge \texttt{Awareness}_i\varphi \rightarrow \texttt{Joy}_i\varphi$$

Proposition 5 (Unsuccessful Trust implies Disappointment)

$$\texttt{Bel}_i\,\texttt{Done}_{j:\alpha}\texttt{Trust}_{i,j}(\alpha,\varphi) \wedge \texttt{Awareness}_i\neg\varphi \rightarrow \texttt{Disappointment}_i\neg\varphi$$

This means that we feel disappointed if what we trusted does not in fact occur. For example, a businessman trusted his partner to arrive on time to negotiate a contract. The businessman feels disappointed if the partner has not yet arrived at the scheduled time.

DisTrust and Fear. Distrust features a negative expectation, involving fear of the other [21, 2]. We state the relation between *Distrust* based on ability and *Fear* as Proposition 6.

Proposition 6 (DisTrust implies Fear)

$$\texttt{C-DisTrust}_{i,j}(\alpha,\varphi) \rightarrow \texttt{Fear}_i\neg\varphi$$

This means that if we distrust someone to do an action to bring us something then we fear that our desire might not be fulfilled. For example, the boss might distrust his assistant with the preparation of a report he needs, and more specifically distrusts him to finish the report by the next morning ($\texttt{DisTrust}_{boss,assistant}(finish, report)$). Therefore, he is fearful that he might miss the report the next morning ($\texttt{Fear}_{boss}\neg report$). This proposition will be proved by applying Lemma 2: if we believe that someone is unable to do an action to bring about something, then we believe that there is at least a future world without the expected result of this action.

Lemma 2

$$\texttt{Bel}_i\,\neg\texttt{After}_{j:\alpha}\varphi \rightarrow \texttt{Possible}_i\texttt{F}\neg\varphi$$

Once we distrust someone to do an action to bring about something, we experience fear. If the results are indeed negative, we feel fear-confirmed (formalized as Proposition 7). If, however the action is in fact successfully performed, we feel relief (formalized as Proposition 8).

Proposition 7 (Confirmation of DisTrust implies Fear-confirmed)

$$\texttt{Bel}_i\,\texttt{Done}_{j:\alpha}\texttt{C-DisTrust}_{i,j}(\alpha,\varphi) \wedge \texttt{Awareness}_i\neg\varphi \rightarrow \texttt{FearConfirmed}_i\neg\varphi$$

If the boss realizes that his assistant really did not finish the report ($\texttt{Bel}_{boss}\,\neg report$), he feels fear-confirmed ($\texttt{FearConfirmed}_{boss}\neg report$). Combining the two Propositions 2 and 7, we arrive at a corollary: when we experience that what we distrusted has now happened, we feel distressed about it.

Corollary 2

$$\texttt{Bel}_i\,\texttt{Done}_{j:\alpha}\texttt{C-DisTrust}_{i,j}(\alpha,\varphi) \wedge \texttt{Awareness}_i\neg\varphi \rightarrow \texttt{Distress}_i\neg\varphi$$

Proposition 8 (Non-confirmation of DisTrust implies Relief)

$$\texttt{Bel}_i\,\texttt{Done}_{j:\alpha}\texttt{C-DisTrust}_{i,j}(\alpha,\varphi) \wedge \texttt{Awareness}_i\varphi \rightarrow \texttt{Relief}_i\varphi$$

If the boss discovers that his assistant did in fact finish the report ($\texttt{Bel}_{boss}\,report$), he feels relieved ($\texttt{Relief}_{boss}report$).

5.2 Behavioral validation

Although the propositions that we proved in the previous section are intuitively plausible, some of them have not yet received behavioral validation from the field of experimental psychology. We decided to collect empirical data concerning three propositions in this article, related to the emotions that follow trust when it is confirmed (Proposition 4), and when it is unconfirmed (Proposition 5); and the emotions that follow distrust, when it is unconfirmed (Proposition 8) [5].

Following the analysis in (Section 4) which argues that trust is the conjunction of the intention, the capacity, and the agreement of trustee, the presence of *Agreement* is intentionally fixed for the future test. We therefore operationalize *Trust* as the conjunction of *Intention* and *Capacity*, and *Distrust* as the three remaining cases. Participants to the survey read 8 different stories, following a $2 \times 2 \times 2$ within-subject design. The variables manipulated in the stories were *Intention* (Yes/No), *Capacity* (Yes/No), and *Outcome* (Success/Failure). As an example, here is the story corresponding to *Intention* = *Yes*, *Capacity* = *Yes*, and *Outcome* = *Success*.

> Mr. Boss is the marketing director of a big company. He needs an important financial report before a meeting tomorrow morning, but he has no time to write it because of other priorities. He asks Mr. Support to prepare it and put it on his desk before tomorrow morning.
>
> - *Mr. Boss believes that Mr. Support has the intention to prepare the report in time.*
> - *Mr. Boss believes that Mr. Support is able to prepare the report in time.*
>
> The morning after, Mr. Boss finds the report on his desk when he arrives. In your opinion, what does he feel?

In the condition *Intention* = *No*, "Mr. Boss believes that Mr. Support has the intention to prepare the report in time" was replaced with "Mr. Boss believes that Mr. Support has no intention to prepare the report in time." In the condition *Capacity* = *No*, "Mr. Boss believes that Mr. Support is able to prepare the report in time" was replaced with "Mr. Boss believes that Mr. Support is unable to prepare the report in time." Finally, in the condition *Outcome* = *Failure*, "Mr. Boss finds the report on his desk when he arrives" was replaced with "Mr. Boss does not find the report on his desk when he arrives."

After reading each story, participants rated the extent to which the main character would feel each of 7 emotions, which included our target emotions, satisfaction, disappointment, and relief; but also some emotions that we included for exploratory purposes, such as anger or thankfulness. Ratings used a 6-point scale anchored at *Not at all* and *Totally*.

A total of 100 participants took part in an online survey. The survey was offered in two languages, French (30% of the final sample) and Vietnamese (70%). Language was entered as a control variable in all statistical analyses, but added only a small overall main effect on participants' responses, and will not be discussed any further.

[5] We could not test Proposition 7 for a linguistic reason: Neither in French nor in Vietnamese (the two languages used in our experiment) could we find an everyday term equivalent to 'fear confirmed'.

	Satisfaction		Relief		Disappointment	
	Trust	Distrust	Trust	Distrust	Trust	Distrust
Success	4.9 (1.5)	4.6 (1.6)	2.8 (1.9)	3.6 (1.9)	1.1 (0.6)	1.3 (0.8)
Failure	1.1 (0.5)	1.4 (1.0)	1.3 (1.0)	1.3 (0.9)	4.6 (1.7)	3.2 (1.4)

Table 1: Mean and standard deviations of affective ratings, as a function of Trust and Outcome.

Descriptive statistics are displayed in Table 1. Participants' responses were analyzed by means of a repeated-measure analysis of variance, aimed at detecting statistically reliable effects of Trust and Outcome on our emotions of interest.

Satisfaction. Unsurprisingly, the analysis of variance detected a huge effect of Outcome, $F(1, 98) = 597$, $p < .001$, accounting for most of the observed variance, $\eta_p^2 = .86$. In other terms, Satisfaction is almost perfectly predicted by Outcome alone. The analysis, however, also detects a comparatively small interaction effect Outcome $\times$ Trust, $F(1, 98) = 8.8$, $p < .01$, $\eta_p^2 = .08$, reflecting the fact that success is even more pleasant in case of trust. Table 1 shows that the biggest score of *Satisfaction* is in the case of Trust follows a Success: $M = 4.9$, $SD < 1.5$. The data are in line with what was expected from Proposition 4.

Relief. The analysis detected main effects of Trust, $F(1, 98) = 19.1$, $p < .001$, $\eta_p^2 = .23$; and Outcome, $F(1, 98) = 127$, $p < .001$, $\eta_p^2 = .80$. However, these main effects were qualified by an interaction effect Trust $\times$ Outcome, $F(1, 98) = 12.3$, $p < .001$, $\eta_p^2 = .31$. Table 1 shows that the score of *Relief* is especially high in the case of Success is obtained despite of Distrust: $M = 3.6$, $SD < 1.9$. This interaction reflects our expectation (Proposition 8).

Disappointment. The analysis detected main effects of Trust, $F(1, 98) = 28.4$, $p < .001$, $\eta_p^2 = .16$; and Outcome, $F(1, 98) = 389$, $p < .001$, $\eta_p^2 = .56$. However, these main effects were qualified by an interaction effect Trust $\times$ Outcome, $F(1, 98) = 44.7$, $p < .001$, $\eta_p^2 = .11$. Table 1 shows that the score of *Disappointment* is especially high in the case of Failure is obtained despite of Trust: $M = 4.6$, $SD < 1.7$. This interaction reflects our expectation (Proposition 5).

6 Conclusion

This paper introduced a logical framework that can represent the cognitive structure of emotions, trust, and the formal relations between them. In other terms, it enables to represent the effect of trust (and distrust) on emotions. Furthermore, this logical framework respects the instantaneity of emotions that previous logics of emotions did not fulfill. Finally, the formal relations between emotion and trust laid bare by the logical framework were subjected to a behavioral validation following the methods of experimental psychology. The success of this behavioral validation gives strong support to our approach, which is shown to capture lay users' intuitions about trust-related emotion.

Although we have added time factor into almost emotional formulas, which enables to eliminate rightly emotion when the relevant event has passed a long time, but it have not yet helped us to represent the nature of continuous intensity of emotions. Additionally, this paper has formalized only the effect of trust/distrust

on emotions but not yet the effect of emotions on trust/distrust. These current limitations are also the potential perspective for our future research.

acknowledgement

This work has been supported by the Agence Nationale de la Recherche (ANR), contract No. ANR-08-CORD-005-1, and by a doctoral scholarship awarded by the University of Toulouse, contract No. 26977-2007.

References

[1] Carole Adam, Andreas Herzig, and Dominique Longin. A logical formalization of the OCC theory of emotions. *Synthese*, 168(2):201–248, 2009.

[2] Philippe Aghion, Yann Algan, Pierre Cahuc, and Andrei Shleifer. Regulation and Distrust*. *The Quarterly Journal of Economics*, 125(3):1015–1049, 08 2010.

[3] H. J. Bryce. Formalizing Civic Engagement: NGOs and the Concepts of Trust, Structure, and Order in the Public Policy Process. In *Workshop on Building Trust Through Civic Engagement and for the International Political Science Association, Section on Governance, conference on Government Crisis in Comparative Perspective*, Seoul, Korea, 2007.

[4] C. Castelfranchi and R. Falcone. Social trust: A cognitive approach. In C. Castelfranchi and Y. H. Tan, editors, *Trust and Deception in Virtual Societies*, pages 55–90. Kluwer Academic Publishers, Dordrecht, 2001.

[5] Cristiano Castelfranchi, Rino Falcone, and Emiliano Lorini. *A Non-reductionist Approach to Trust*, pages 45–72. Springer London, London, 2009.

[6] B. F. Chellas. *Modal Logic: an Introduction*. Cambridge University Press, Cambridge, 1980.

[7] E. M. Clarke, E. A. Emerson, and A. P. Sistla. Automatic verification of finite-state concurrent systems using temporal logic specifications. *ACM Transactions on Programming Languages and Systems*, 8(2):244–263, April 1986.

[8] Philip R. Cohen and Hector J. Levesque. Intention is choice with commitment. *Artificial Intelligence Journal*, 42(2–3):213–261, 1990.

[9] Ronald de Sousa. *The Rationality of Emotion*. MIT Press, 6th edition, 2001.

[10] Fred Dretske. *Naturalizing the Mind*. MIT Press, 1995.

[11] Michael J. Fischer and Richard E. Ladner. Propositional dynamic logic of regular programs. *Journal of Computer and System Sciences*, 18(2):194–211, 1979.

[12] N. H. Frijda. *The Emotions*. Cambridge University Press, 1986.

[13] Benoit Gaudou, Andreas Herzig, and Dominique Longin. A Logical Framework for Grounding-based Dialogue Analysis. *Electronic Notes in Theoretical Computer Science (ENTCS)*, 157(4):117–137, 2006.

[14] Paul Gochet and Pascal Gribomont. Epistemic Logic. In Dov Gabbay and John Woods, editors, *Twentieth Century Modalities*, volume 7 of *Handbook of the History of Logic*, pages 99–195. Elsevier, amsterdam edition, 2006.

[15] D. Harel, D. Kozen, and J. Tiuryn. *Dynamic Logic*. MIT Press, Cambridge, 2000.

[16] Andreas Herzig and Dominique Longin. C&L intention revisited. In Didier Dubois, Chris Welty, and Mary-Anne Williams, editors, *Proc. 9th Int. Conf. on Principles of Knowledge Representation and Reasoning (KR 2004), Whistler, Canada, June 2–5*, pages 527–535. AAAI Press, 2004.

[17] Andreas Herzig, Emiliano Lorini, Jomi F. Hübner, Jonathan Ben-Naim, Olivier Boissier, Cristiano Castelfranchi, Robert Demolombe, Dominique Longin, Laurent Perrussel, and Laurent Vercouter. Prolegomena for a logic of trust and reputation. In Guido Boella, Gabriella Pigozzi, Munindar Singh, and Harko Verhagen, editors, *International Workshop on Normative Multiagent Systems (NorMAS), Luxembourg, 15/07/2008-16/07/2008*, pages 143–157. University of Luxembourg Press, 2008. ISBN: 2919940481.

[18] J. Hintikka. *Knowledge and Belief: An Introduction to the Logic of the Two Notions*. Cornell University Press, Ithaca, 1962.

[19] Bernd Lahno. On the emotional character of trust. *Ethical Theory and Moral Practice*, 4(2):171–189, Jun 2001.

[20] Richard S. Lazarus. *Emotion and Adaptation*. Oxford University Press, 1991.

[21] R. J. Lewicki and C. Wiethoff. Trust, trust development, and trust repair. In M. Deutsch and P. T. Coleman, editors, *The handbook of conflict resolution: Theory and practice*, pages 86–107. Jossey-Bass, 2000.

[22] Andrew Ortony, G.L. Clore, and A. Collins. *The cognitive structure of emotions*. Cambridge University Press, Cambridge, MA, 1988.

[23] A. N. PRIOR. *Time and Modality*. Greenwood Press, 1955.

[24] Rainer Reisenzein. Emotions as metarepresentational states of mind: Naturalizing the belief-desire theory of emotion. *Cognitive Systems Research*, 10:6–, 03 2009.

[25] K. R. Scherer. *Appraisal Processes in Emotion : Theory, Methods, Research*, chapter Appraisal Considered as a Process of Multilevel Sequential Checking, pages 92–120. Oxford University Press, New York, 2001.

[26] K. R. Scherer, A. Schorr, and T. Johnstone, editors. *Appraisal Processes in Emotion: Theory, methode, Research*. Affective Science. Oxford university Press, 2001.

[27] B. R. Steunebrink, M. Dastani, and J. J. Ch. Meyer. A logic of emotions for intelligent agents. In *Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence (AAAI'07)*, pages 142–147. AAAI Press, 2007.

[28] B.R. Steunebrink, Mehdi Dastani, and John-jules Meyer. Towards a quantitative model of emotions for intelligent agents. In Reichardt and Levi, editors, *Proceedings of the 2nd Workshop on Emotion and Computing - Current Research and Future Impact*, Osnabrück, Germany, 2007.

[29] D. N. Walton and E. C. Krabbe. *Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning*. State University of New-York Press, NY, 1995.

Chapter 5

New research perspectives

5.1 Introduction

On one hand, there is a lot of formal models of cognitive agents but there are few fully implemented systems because of too strong hypothesis (that do not work with respect to real applications) and of efficiency reasons. On the other hand, there are simulation platforms for multi-agent systems aiming to simulate scenarios with a great number of agents. But in this latter case, agents are often simple and the simulation cannot explain individually why a particular agent has made some action or other. (Ideally, individual actions should be able to explain the collective action.)

Whereas cognitive agent systems use generally a restricted number of agents (modeling oriented on individual behavior), simulation platforms focus on the global behavior of the system. The objective of this program of research is to reconciliate these two approaches.

Our approach has two steps: the development and the implementation of a simplified formal model of cognitive agent, and the implementation of this model in a simulation platform. We will implement an example that will be used as a test of our model. This example is about the evacuation of a crowd out of a public area where a crisis situation has happened. The GAMA platform of simulation will provide us a good tool for the management of spatiotemporal aspects.

5.2 Génèse

Comme le montre le détail de nos activités, nous avons depuis le début de notre carrière travaillé sur des modèles formels d'agent cognitif. Fin 2006, nous avons répondu à un appel d'offre du PRES de Toulouse qui offrait des bourses du ministère pour des sujets de thèses co-encadrées par des personnes provenant d'au moins deux établissements différents de la région toulousaine. Le sujet, déposé en collaboration avec Jean-François Bonnefon[1] portait sur la relation entre confiance et émotion et contenait à la fois des aspects formels et des aspects empiriques. C'était à la fois une façon de lier deux de nos domaines de recherche du moment (la confiance et l'émotion) et d'effectuer un retour vers des travaux pluridisciplinaires tels que nous les menions au début des années 2000. Après recherche d'un candidat, nous avons recruté Manh Hung NGUYEN, un étudiant vietnamien. L'année suivante, un autre étudiant (Benoit GAUDOU) que nous co-encadrions avec Andreas Herzig sur la notion de grounding partait en post-doctorat à l'Institut de la Francophonie en Informatique (IFI) à Hanoï pour travailler avec Alexis Drogoul, directeur de recherche à l'IRD, sur la plateforme multi-agents GAMA que ce dernier développe à l'IFI depuis quelques années. M. NGUYEN venait lui-même de l'IFI et nous avons par ce biais amorcé une collaboration informelle avec le professeur Tuong Vinh HO. L'idée a peu à peu germé d'essayer de concilier des modèles formels du raisonnement avec des modèles implémentés de systèmes multi-agents et a été concrétisée en 2013 avec l'obtention d'une bourse de l'Université des Sciences et Technologie de Hanoï (avec des avantages en nature du ministère français des affaires étrangères). Ces travaux de thèse sont actuellement menés par un étudiant vietnamien (Xuan Hien TA). Parallèlement à cela, nous avons co-encadré un troisième étudiant vietnamien de l'IFI (Van Tho NGUYEN) dans le cadre de son TPE de master 1 et avec lequel nous avons publié en 2014 un article sur une simulation d'évacuation d'un lieu public. Il nous semble donc naturel de continuer dans cette voie puisque ces thématiques de recherche s'inscrivent dans nos préoccupations actuelles et que nos collaborations avec le Vietnam offrent un terrain propice à des développements concrets de nos travaux.

[1] Chercheur CNRS en psychologie cognitive au laboratoire CLLE-LTC de l'Université Toulouse 2 Jean-Jaurès.

5.3 Cadre sociétal et enjeux

Ces dernières années, il est aisé de constater que les nouvelles technologies envahissent de plus en plus notre quotidien, que ce soit dans le cadre d'applications ludiques, sociales, professionnelles, etc. Devant l'importance du phénomène et face à des demandes de plus en plus fines et nombreuses, les services augmentent en même temps que leur propre complexité dans le but de fournir une réponse de plus en plus proche de ce que l'utilisateur attend tout en maximisant son accessibilité. Cela implique que ces systèmes soient crédibles aux yeux des utilisateurs et disposent dans ce but de capacités d'interaction évoluées, ce qui passe immanquablement par une capacité à comprendre, raisonner sur et prédire les états mentaux en général des utilisateurs et leurs émotions en particulier (cet aspect-là étant bien plus récent). Certains poussent cette crédibilité jusqu'à exiger d'elle que les agents ne soient pas seulement honnêtes et fiables, mais fournissent l'illusion de la vie [14].

Parallèlement à ce domaine où les recherches se focalisent sur des agents cognitifs de plus en plus évolués, le domaine des simulations en tout genre à base d'agents relativement peu évolués a également pris beaucoup d'ampleur. Que ce soit la simulation de phénomènes physiques, chimiques ou biologiques, des besoins ont aussi émergé dans le domaine de la gestion de crise à laquelle nous nous sommes intéressés pour des raisons historiques (comme expliqué dans la section précédente).

L'enjeu est de taille : il s'agit de comprendre comment améliorer la survie des personnes en cas de crise. Ce domaine intéresse aujourd'hui un certain nombre de chercheurs en France et ailleurs dans le monde, et concerne des domaines d'application où la vie de personnes est mise en péril suite à des événements aussi divers que des accidents de la route à grande échelle, des épidémies mortelles en tout genre, des catastrophes chimiques, etc., en passant par toute sorte de catastrophes naturelles. Ce dernier type d'événement est au centre de l'intérêt de pays comme le Vietnam où les catastrophes climatiques sont régulièrement à l'origine de drames, le niveau économique du pays ne permettant pas, comme c'est le cas au Japon avec la construction de maisons parasismiques par exemple, de prémunir la population à grande échelle.

Dans cet optique, l'objectif de ces simulations est double : tenter de reproduire les situations passées et comprendre ainsi comment les choses se sont passées, et proposer ensuite des améliorations au niveau de la prévention,

de l'évacuation des lieux, de l'intervention des secours, etc. afin de réduire le nombre de victime lors d'une prochaine catastrophe.

Ainsi, l'informatique se doit aujourd'hui de fournir des modèles tant pour des applications à base d'agents cognitifs ayant des caractéristiques les plus proches possibles de celles des humains (afin d'obtenir une interaction la plus naturelle possible), que pour des applications intégrant potentiellement un grand nombre d'agents généralement peu évolués mais dont le comportement global doit pouvoir être décrit au sein d'un mouvement de foule et correspondre à des cas concrets (et réels) de situations.

Un des enjeux de notre programme est de répondre à cette double demande en proposant un modèle d'agent cognitif doté de capacités de raisonnement simplifiées mais capable d'expliquer son comportement (capturé au travers de ses actions) ainsi que les facteurs (notamment émotionnels, mais pas seulement) ayant influencé ses choix. Cette modélisation sera ensuite utilisée au sein d'une architecture multi-agents afin de simuler l'évacuation d'un lieu public en situation de crise.

5.4 Cadre scientifique

Les modèles formels d'agent, dont on peut trouver les prémisses fin des années 70 début des années 80 avec les travaux d'Allen, Cohen et Perrault sur l'action et le langage (voir [6] par exemple) ont abouti une dizaine d'année plus tard aux travaux fondateurs de Cohen et Levesque sur une logique des états mentaux (croyance, but et intention) et de l'action [32, 33]. Ces modèles ont ensuite été précisés tant au niveau des concepts que de leur formalisation (voir [130, 124, 160, 83] par exemple) et ont été étendus, par exemple à des modalités déontiques [23] ou sociales (croyance mutuelle, croyance partagée, croyance de groupe, acceptance, etc.) en même temps que naissaient des implémentations plus ou moins complètes de ces systèmes (voir [74] ou [114] par exemple). Plus récemment sont apparues des extensions aux émotions [113] auxquelles nous avons nous-mêmes contribué [3].

Nombreux sont les travaux en psychologie ayant montré le rôle central de l'émotion dans la cognition et l'interaction sociale [117] ainsi qu'à chaque étape du processus de raisonnement. Du point de vue de l'informatique le défi semble grand et l'exemple de la chambre chinoise élaboré par Searle pour montrer le fossé infranchissable selon lui entre le fonctionnement des ordina-

teurs et celui de notre cerveau n'est point pour rassurer.[2] Mais comme le disent Ortony, Clore et Collins [115, p. 17] "le but (...) n'est pas de créer des machines dotées d'émotions, mais de créer des modèles informatiques pouvant comprendre quelles émotions les gens peuvent éprouver, et sous quelles conditions. De tels systèmes seraient ainsi capables de prédire et d'expliquer les émotions humaines, pas de les ressentir."

Depuis plusieurs années nous travaillons sur des modèles formels des émotions. Si l'association des termes *modèle formel* et *émotion* peut paraître étonnante, voire antinomique, elle n'en est pourtant pas moins naturelle. Cette question précisément a fait l'objet d'un débat animé en psychologie qui a débuté dans les années 80 [161] et dont les principaux protagonistes ont été Zajong et Lazarus. La question était de savoir si l'émotion était dans la cognition ou si c'était un phénomène indépendant qui tendait au contraire (selon l'opinion populaire en provenance directe de Platon) à s'opposer à la cognition, voire à la précéder. Lazarus estime pour sa part que émotion, cognition, et motivation sont "plus ou moins des fictions de l'analyse scientifique, dont l'indépendance n'existe pas réellement dans la nature" [95]. L'émotion ne peut se manifester sans la présence *simultanée* de processus cognitifs et motivationnels (potentiellement inconscients). L'émotion est une réponse à une certaine configuration de croyances et de buts, condition nécessaire pour que cette émotion soit pertinente par rapport à notre état mental. Ainsi, bien qu'une pensée puisse apparaître sans émotion associée l'inverse n'est pas vrai : une émotion ne peut se manifester complètement détachée de toute pensée (elle a nécessairement un contenu). Néanmoins il faut considérer que les émotions et les pensées se produisent de manière permanente et que l'une peut tout à fait être à l'origine de l'autre. La dimension cognitive de l'émotion décrivant notamment ses conditions de déclenchement (appelée généralement "structure cognitive de l'émotion", voir par exemple [132]) fait appel aux croyances et aux buts de l'individu[3] ce qui fait des logiques des états mentaux un outil particulièrement adapté pour les représenter.

Du point de vue de l'implémentation, il existe des langages d'agent basés

[2]Dans [139, p. 42] Searle illustre le fonctionnement interne d'un ordinateur qui parlerait le Chinois en montrant que ce fonctionnement ne requiert ni n'induit pour l'ordinateur aucune capacité à "comprendre" les phrases qu'il "prononce".

[3]Les buts sont ici à prendre au sens large : ils incluent aussi bien ceux propres à satisfaire nos désirs que ceux propres à satisfaire nos obligations ou nos normes morales internalisées (w.r.t. les normes dans lesquelles nous nous reconnaissons et que nous nous fixons pour but de satisfaire).

sur des architectures comme Golog qui est apparu à la fin des années 90 [74] ou des architectures BDI comme Jason [19] (qui est un interpréteur d'une version étendue de AgentSpeak [121]) mais aucun ne gère l'émotion comme composant de la prise de décision. Pour l'immense majorité des modèles logiques, soit ceux-ci ne sont pas implémentés, soit il existe un démonstrateur de théorème permettant d'évaluer si une formule est valide ou non mais ne constituant pas en soi un système multi-agent capable de déterminer quelle formule il faut démontrer. Par ailleurs, souvent ces logiques ont des hypothèses très fortes communes à tous les agents (connaissance du monde, connaissance des actions des autres, omniscience, etc.). De plus, la complexité des logiques à implémenter rend toute implémentation gourmande en espace mémoire et en temps de calcul alors que dans le même temps ces logiques permettent de déduire certains faits non réalistement accessibles à notre cognition. Par exemple, l'imbrication d'un nombre potentiellement très important d'opérateurs (l'agent i croit que le but de l'agent j est que k croit que l ait pour but que...) est un facteur de complexité inutile puisqu'au delà de deux ou trois imbrications tout individu a du mal à se représenter la signification d'une telle imbrication.

Alors que se développaient en psychologie des études sur les situations d'urgence tendant à montrer que l'émotion joue un rôle prépondérant[4] (voir par exemple [125]) l'émotion a également fait son apparition dans le domaine des simulations multi-agents de gestion de crise (voir [149, 97] par exemple). Dans ces systèmes, à l'inverse de ce qui est fait dans les modèles logiques de l'émotion, celle-ci est souvent représentée *via* un simple label auquel il est attaché une valeur numérique entre 0 et 1. Quand cette valeur est booléenne elle représente le fait que l'émotion soit active ou non et quand elle est un nombre décimal elle représente son intensité. [5]

Il est intéressant de remarquer qu'en logique, on ne représente souvent que la structure cognitive de l'émotion alors que dans les simulations on ne s'attache qu'à son intensité (qui indique l'importance de l'émotion dans la prise de décision et influence les actions de l'agent). Pourtant il semble que les deux composantes soient comme les deux faces d'une même pièce.

[4]Tout en dénonçant le mythe de la panique massive, largement utilisée dans les films catastrophe mais relativement absente des cas réels [44, 120].

[5]Il existe des modèles qui ne sont pas basés sur des agents individuels mais sur d'autres concepts tels que la mécanique des fluides, les forces sociales ou autres, mais ces approches sortent du cadre de notre travail et ne sont propres qu'à modéliser un phénomène, non à l'expliquer.

Ainsi, une émotion peut se déclencher au sens où les conditions sur les états mentaux d'un agent sont satisfaites sans pour autant ressentir une émotions particulière. Par exemple, nous ne sommes pas tristes en permanence pour toutes les choses que nous aimerions voir vraies mais qui ne le sont pas : notre cognition agit en fait comme un filtre qui évacue toutes les émotions que nous pourrions ressentir mais auxquelles nous n'attachons pas d'importance (w.r.t. dont l'intensité est inférieure à un certain seuil qu'on pourrait décrire comme un seuil d'activation à la Anderson [10]). Par ailleurs l'intensité modifie les tendances à l'action en opérant une restriction de celles issues du déclenchement de l'émotion proprement dite : quand nous avons très peur par exemple, nous avons tendance à marcher plus vite que lorsque notre niveau de peur est moindre, ou à nous sauver alors que c'est une option qui n'est pas privilégiée dans le cas d'une petite frayeur (voir le facteur d'agitation décrit dans [17] par exemple). Comme le montrent Ortony, Clore et Collins [115] l'intensité ne résulte pas seulement de la force de la croyance et de l'importance du but : elle dépend aussi de certains facteurs externes comme la fréquence à laquelle on est confronté à la situation génératrice d'émotion, la date du dernier ressenti de cette émotion, les facteurs influençant notre santé, voire notre vie (on ne sera pas effrayé de marcher sur un fil si celui-ci est situé à 10cm du sol mais on pourrait l'être s'il est situé à 10m par exemple), etc. En revanche, structure cognitive et intensité ne sont pas totalement déconnectées dans le sens où une intensité non nulle nécessite un déclenchement préalable de l'émotion correspondante et où le déclenchement d'une émotion est nécessairement lié à un certain degré d'intensité, même si celui-ci n'est pas perceptible pour l'agent.

D'une manière générale les simulations multi-agents actuelles sont souvent basées sur un (grand) nombre de variables dont les critères d'évolution (souvent non déterministes) empêchent d'identifier clairement les propriétés du système. Par exemple les conditions de déclenchement des émotions associées à la structure cognitive de l'émotion ne sont pas explicitement représentées et sont (plus ou moins) diluées dans une mécanique parfois complexe de variables, de même que les comportements associés. C'est également le cas pour les autres états mentaux et leurs propriétés qui ne sont pas représentés au sein d'une théorie claire aux propriétés correctement identifiées. Ainsi, si un agent i perçoit qu'un agent j a très peur, l'approche classique d'une simulation consistera par exemple à déterminer aléatoirement si i est influencé ou non par j et si c'est le cas, dans quelle mesure (en fonction d'un autre paramètre propre à l'agent). Ce faisant on est donc capable de modéliser

l'influence de l'émotion de j sur l'agent i mais on n'explique pas pourquoi cette influence a lieu du point de vue de la cognition des agents. Nous ne prétendons pas qu'il serait impossible aux simulations de développer des mécanismes explicatifs mais seulement que ce n'est pas ce qui intéressent les personnes construisant ces modélisations.

5.5 Programme de recherche et contributions à l'état de l'art

La contribution que nous pensons pouvoir apporter se situe à la limite des systèmes multi-agents et de l'intelligence artificielle. Notre but est de développer une simulation multi-agent de l'évacuation d'un lieu public en situation d'urgence fondée sur une théorie logique simplifiée mais implémentée. Comme nous l'avons déjà dit plus haut le but d'une telle simulation est non seulement de reproduire une catastrophe passée mais également d'aider à déterminer comment le nombre de victimes aurait pu être réduit. La finesse du comportement des agents au sein de la simulation est ainsi particulièrement important (pour pouvoir représenter des situations réelles) et il s'agit de faire profiter ce domaine des connaissances importantes acquises dans le domaine de la formalisation des états mentaux et des actions des agents, y compris la composante émotionnelle.

Nous souhaitons donc développer en premier lieu une implémentation correspondant à un modèle simplifié d'architecture BDI mais dont les propriétés logiques peuvent être étudiées. L'idée est de représenter les croyances et les buts des agents par des vecteurs d'information trivalués. Le terme "buts" est ici à interpréter au sens large : cela concerne à la fois des buts à accomplir ainsi que des buts à maintenir, et ces buts peuvent être issus soit de désirs soit de normes morales internalisées que l'agent se fixe de respecter. Traditionnellement dans les logiques doxastiques un opérateur de croyance peut engendrer trois états (la croyance qu'une certaine proposition est vraie, celle qu'elle est fausse, et un état où l'agent ne croit ni que cette proposition est vraie ni qu'elle est fausse). C'est pour rendre compte de ces trois états que nous choisissons des informations à trois valeurs (et non à deux comme c'est traditionnellement le cas en logique). Un vecteur d'information particulier représentera l'état réel du monde auquel chaque agent n'a que partiellement accès et sur lequel il ne peut intervenir que de manière par-

tielle. (En d'autres termes, un agent ne connaît pas l'état complet du monde et il ne peut intégralement le changer ; les propositions auxquelles il a accès dans le monde ne correspondent pas nécessairement à celle dont il peut modifier la valeur.) Finalement, le monde réel ainsi que les croyances et les buts des agents sont soumis à des contraintes d'intégrité pour exprimer par exemple que lorsqu'un feu de signalisation routière fonctionne, l'ampoule rouge est allumée, ou l'ampoule orange est allumée, ou la verte, mais pas les trois à la fois. Nous avons contribué à formaliser la structure cognitive d'un certain nombre d'émotions [3] mais nous n'avons jamais (comme c'est souvent le cas dans les modèles logiques) décrit leur intensité. À la lumière des nombreux facteurs pouvant influencer cette intensité nous prévoyons de la représenter par une valeur indépendante de la structure cognitive correspondante mais déterminée à partir des croyances sur certains faits du monde qui dépendent de l'application modélisée. Cette valeur sera dans la mesure du possible qualitative. Globalement, l'intensité des émotions sera un champs nouveau d'investigation pour nous et l'occasion de proposer un modèle formel de l'intensité des émotions.

Le système visé devra évoluer dynamiquement à plusieurs niveaux. Au niveau du monde réel, cette évolution se fera *via* les actions des agents. Cela pose le problème de l'agrégation des effets : que se passe-t-il si les actions sont contradictoires ? Est-ce qu'une majorité d'agents suffit à imposer le résultat d'une action ou non ? Que se passe-t-il pour les agents qui, se basant sur leurs croyances, pensent accomplir une action dont le résultat escompté correspond déjà à l'état actuel du monde ? Que se passe-t-il si le résultat des actions ne respecte pas les contraintes d'intégrité ? Ce sont autant de questions auxquelles il faudra répondre et qui permettront d'étudier des solutions différentes afin de choisir la plus adaptée à une situation donnée.

Au niveau des croyances des agents la dynamique est double. D'une part le monde réel étant partiellement observable par les agents, un changement dans le monde réel (comme résultat des actions des agents) peut avoir pour conséquence d'influencer les croyances des agents pour qui la propriété du monde réel qui a changé est observable. D'autre part on souhaite que la croyance d'un agent à propos d'une certaine proposition donnée (mais ne correspondant pas à un fait observable du monde) puisse être influencée par l'opinion de certains autres agents. C'est un domaine qui a largement été étudié depuis une trentaine d'années (voir les travaux fondateurs d'Alchourrón, Gärdenfors et Makinson [5] sur la révision et de Katsuno et Mendelson [90] sur la mise à jour). Mais d'une manière générale différentes

manières d'agréger les opinions existent et l'un des objectifs de ce programme sera d'étudier, dans le cadre applicatif qui est le nôtre, celui qui semble le plus réaliste.

Au niveau des émotions des agents la dynamique va se dérouler à deux niveaux. Le premier concerne la structure cognitive des émotions qui décrit leurs conditions de déclenchement au sein de la cognition d'un agent (par exemple la structure cognitive de la tristesse est typiquement une incongruence entre les croyances et les désirs d'un agent). Du fait que ces conditions de déclenchement sont traduites en termes de croyances et de buts, tant un changement au niveau du monde réel que l'influence des opinions des autres peut avoir pour effet de changer les croyances d'un agent, donc de déclencher en lui une émotion. Le second niveau concerne l'intensité des émotions. Face à certaines croyances que l'agent acquiert l'intensité des émotions varie. Par exemple, l'approche d'un danger va augmenter l'intensité de la peur et la présence d'une personne apte à nous protéger ou nous rassurer va la faire diminuer. Le temps est également un facteur naturel de diminution de l'intensité des émotions [115]. Parallèlement à cela des travaux en psychologie montrent que l'émotion se diffuse au sein d'une foule (voir [73, 72] par exemple). Cela signifie que de voir quelqu'un qui a peur ou qui est joyeux va nous pousser à avoir peur ou à être joyeux à notre tour. Ce phénomène est appelé *contagion émotionnelle* dans la littérature. (Il ne doit pas être confondu avec l'empathie qui fait référence à un mécanisme où un agent, se "mettant à la place" de quelqu'un, se met à ressentir ce que cette personne ressent.) Dans [73] les auteurs avancent que c'est le niveau d'intensité le plus élevé parmi ceux attachés aux agents constituant notre voisinage qui nous impacte. Mais nous ne sommes pas tous impactés de la même manière et si l'intensité la plus élevée est x cela ne signifie pas pour autant que l'intensité de notre émotion sera x (ni même que l'on éprouvera une émotion !). Là encore nous prévoyons de mener des investigations en nous basant sur les travaux les plus récents dans le domaine.

Dans un second temps, en nous appuyant sur le modèle et l'implémentation précédents, nous souhaitons utiliser comme cadre applicatif l'évacuation d'un lieu public en situation d'urgence : des agents sont répartis dans un lieu public (par exemple un magasin) contenant des obstacles (les rayons contenant les produits en vente par exemple). Un feu se produit à un endroit donné et se propage. Les agents évacuent les lieux en se dirigeant vers les issues de secours qu'ils connaissent ou qu'ils voient, ou vers l'issue par laquelle ils sont rentrés. Cette évacuation se fait en évitant les

obstacles et les autres agents. Ils peuvent éventuellement repérer un agent de sécurité afin de le suivre (ce qui est un facteur de diminution du stress). Les agents tendent également à se regrouper par connaissance (familles ou amis par exemple) ce qui est également un facteur de diminution du stress. Au début, tous les agents ne perçoivent pas nécessairement tous le feu. Le stress de ceux qui sont au courant se diffuse à leurs voisins. Selon la vitesse de propagation du feu et le temps mis par les agents à évacuer les lieux, des agents peuvent rester prisonniers des flammes. Le feu est ici une manifestation possible de crise, mais cela peut aussi bien être un tremblement de terre (qui se manifestera par des chutes d'objets à différents endroits du magasin et de manière simultanée) ou autre chose encore (inondation, effondrement d'une partie du bâtiment, etc.).

Ces recherches devraient permettre de contribuer à l'état de l'art de la manière suivante :

- La production d'un modèle théorique d'architecture BDI simplifiée : celle-ci pourra inclure différents mécanismes plus ou moins évolués pour gérer la dynamique des attitudes mentales (observation du monde réel, influence des autres agents, évolution des buts en fonction de l'état du monde, etc.), voire de la situation spatiale de l'agent.
- Un modèle théorique de la formalisation de l'intensité des émotions et la variation de cette intensité au cours du temps.
- Un modèle théorique du processus de contagion émotionnelle décrivant comment un agent est influencé par les émotions des autres agents.
- Un modèle théorique de la gestion de la dynamique du monde réel, des croyances, des buts et des émotions (structure cognitive et intensité).
- L'implémentation complète de cette architecture BDI simplifiée dans un langage de type JAVA.
- L'intégration de cette architecture à une plate-forme de simulation multi-agents (GAMA).
- Le développement d'une application de simulation de l'évacuation d'un lieu public en situation de crise.

5.6 Conclusions

Du point de vue du déroulement de notre programme de recherche, nous pensons mener conjointement la formalisation de ce cadre et son implémentation. L'aspect implémentation est particulièrement important pour nous pour plusieurs raisons : tout d'abord, ce sera l'occasion de tester notre modèle théorique, celui-ci étant de fait relativement simple pour rendre cette implémentation possible. Le cadre sur lequel sera fondé l'implémentation étant relativement bien établi, nous devrions pouvoir déduire certaines propriété formelles du système (par exemple, dans quelles conditions arriverons-nous à une situation où le système est stable ?).

Un autre intérêt de l'implémentation serait de fournir à notre équipe un outil suffisamment modulable pour implémenter, même de façon partielle, une partie des théories formelles qui y sont développées. L'idée sous-jacente est que tant que nous restons dans un cadre logique, nous conservons la possibilité d'utiliser ponctuellement des démonstrateurs de théorèmes externes, ce qui est impossible si l'agent n'a pas été modélisé au sein d'une théorie clairement identifiée.

Dans un second temps, nous pensons qu'appliquer cette architecture à un cas concret tel que l'évacuation d'un lieu public en situation d'urgence permettra d'instancier un certain nombre de choix qui risquent de rester purement arbitraires au niveau de l'architecture elle-même (notamment les procédures d'agrégation des croyances ou le processus de contagion émotionnelle) : le cadre imposé par la simulation d'une situation donnée guidera très certainement nos choix. Par ailleurs, la thématique de l'évacuation en situation d'urgence nous a été fournie naturellement pour des raisons expliquées dans la section concernant la genèse de ce programme de recherche. Nous espérons que cela sera l'occasion pour nous de développer plus en avant une collaboration avec le Vietnam pour qui, comme nous l'avons expliqué plus haut, le développement de ce type de simulation est particulièrement important.

Bibliography

[1] C. Adam and D. Longin. La honte: quand émotion et raisonnement sont liés. *Revue d'Intelligence Artificielle*, 28(1):43–66, 2014.

[2] Carole Adam, Benoit Gaudou, Dominique Longin, and Emiliano Lorini. Logical modeling of emotions for Ambient Intelligence. In Fulvio Mastrogiovanni and Nak-Young Chong, editors, *Handbook of Research on Ambient Intelligence and Smart Environments: Trends and Perspectives*, pages 108–127. IGI Global, http://www.igi-global.com, 2011.

[3] Carole Adam, Andreas Herzig, and Dominique Longin. A logical formalization of the OCC theory of emotions. *Synthese*, 168(2):201–248, 2009.

[4] G. Aist, B. Kort, R. Reilly, J. Mostow, and R.W. Picard. Experimentally augmenting an intelligent tutoring system with human-supplied capabilities: Adding human-provided emotional scaffolding to an automated reading tutor that listens. In *Proceedings of the Fourth IEEE International Conference on Multimodal Interfaces (ICMI 2002)*, pages 483–490. IEEE Computer Society, 2002.

[5] Carlos E. Alchourrón, Peter Gärdenfors, and David Makinson. On the logic of theory change: Partial meet contraction and revision functions. *The Journal of Symbolic Logic*, 50(2):510–530, June 1985.

[6] J. F. Allen and R. Perrault. Analyzing intention in utterances. *Artificial Intelligence*, 15:143–178, 1980.

[7] L. Amgoud and I. Rahwan. An Argumentation-based Approach for Practical Reasoning. In G. Weiss and P. Stone, editors, *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2006)*, pages 347–354. ACM Press, 2006.

[8] A. Anderson. A reduction of deontic logic to alethic modal logic. *Mind*, 22:100–103, 1958.

[9] J. R. Anderson. *The architecture of cognition*. Harvard University Press, Cambridge, MA, 1983.

[10] J. R. Anderson. *The Adaptive Character of Thought*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1990.

[11] J. R. Anderson. *Rules of the Mind*. Lawrence Erlbaum Associates, NJ, 1993.

[12] G. E. M. Anscombe. *Intention*. Harvard University Press, Cambridge, Massachusetts, 2nd edition, 1963.

[13] John L. Austin. *How To Do Things With Words*. Harvard University Press, 2nd edition, 1962.

[14] J. Bates. The role of emotion in believable agents. *Communications of the ACM*, 37(7), 1994.

[15] N. Belnap, M. Perloff, and M. Xu. *Facing the future: agents and choices in our indeterminist world*. Oxford University Press, New York, 2001.

[16] Jonathan Ben-Naim, Dominique Longin, and Emiliano Lorini. Formalisation de systèmes d'agent cognitif, de la confiance, et des émotions. In Pierre Marquis, Odile Papini, and Henri Prade, editors, *Représentation des connaissances et formalisation des raisonnements*, volume 1 of *Panorama actuel de l'IA : ses bases méthodologiques, ses développements*, chapter 16, pages 1–24. Cépaduès, http://www.cepadues.com, 2014.

[17] V.A. Bondarev and O.M. Bondareva. Psychological features of seamen's activity in emergency situations. *TransNav - International Journal on Marine Navigation and Safety of Sea Transportation*, 4(4):453–457, 2010.

[18] Jean-François Bonnefon, Dominique Longin, and Manh Hung Nguyen. A Logical Framework for Trust-Related Emotions. *Electronic Communications of the EASST, Formal Methods for Interactive Systems 2009*, 22:1–16, 2009.

[19] Rafael H. Bordini, Jomi Fred Hübner, and Michael Wooldridge. *Programming Multi-agent Systems in AgentSpeak Using Jason*. Wiley-Blackwell, 2007.

[20] M. Bratman. *Intentions, plans, and practical reason*. Harvard University Press, Cambridge, 1987.

[21] M. E. Bratman. Practical reasoning and acceptance in context. *Mind*, 101(401):1–15, 1992.

[22] Michael E. Bratman. What is intention? In Philip R. Cohen, Jerry Morgan, and Martha E. Pollack, editors, *Intentions in Communication*, chapter 2, pages 15–31. MIT Press, Cambridge, MA, 1990.

[23] Jan Broersen, Mehdi Dastani, Joris Hulstijn, Zisheng Huang, and Leendert van der Torre. BOID architecture: Conficts between beliefs, obligations, intentions and desires. In *Proceedings of the fifth international conference on Autonomous agents (AGENTS'01)*, pages 9–16, NY, USA, 2001. ACM Press.

[24] C. Castelfranchi and I. Poggi. Blushing as a discourse: Was darwin wrong? In W.R̃ay Crozier, editor, *Shyness and Embarrassment*, chapter 8, pages 230–251. Cambridge University Press, 1990.

[25] C. Castelfranchi and Y. H. Tan, editors. *Trust and Deception in Virtual Societies*. Kluwer Academic Publishers, Dordrecht, 2001.

[26] H.-N. Casteñada. *Thinking and doing*. D. Riedel, Dordrecht, Holland, 1975.

[27] Marcos Alexandre Castilho, Olivier Gasquet, and Andreas Herzig. Formalizing action and change in modal logic I: the Frame Problem . *Journal of Logic and Computation*, 9(5):701–735, 1999.

[28] Maud Champagne and Dominique Longin. Non literal communication: From pragmatical to logical and psycholinguistical aspects. Seventh Int. Colloquium on Cognitive Sciences (ICCS'01), Donostia–San Sabastian, Spain, May 9–12, 2001.

[29] B. F. Chellas. *Modal Logic: an Introduction*. Cambridge University Press, Cambridge, 1980.

[30] R. M. Chisholm. Freedom and action. In K. Lehrer, editor, *Freedom and Determinism*. Random House, New York, 1966.

[31] L. J. Cohen. *An essay on belief and acceptance*. Oxford University Press, New York, USA, 1992.

[32] Philip R. Cohen and Hector J. Levesque. Intention is choice with commitment. *Artificial Intelligence Journal*, 42(2–3):213–261, 1990.

[33] Philip R. Cohen and Hector J. Levesque. Persistence, intentions, and commitment. In Philip R. Cohen, Jerry Morgan, and Martha E. Pollack, editors, *Intentions in Communication*. MIT Press, 1990.

[34] Philip R. Cohen, Jerry Morgan, and Martha E. Pollack, editors. *Intentions in Communication*. MIT Press, Cambridge, MA, 1990.

[35] Marco Colombetti and Mario Verdicchio. An Analysis of Agent Speech Acts as Institutional Actions. In Cristiano Castelfranchi and Lewis W. Johnson, editors, *Proceedings First Int. Joint Conf. on Autonomous Agents and MultiAgent Systems (AAMAS-2002)*, volume 3, pages 1157–1164. ACM Press, 2002.

[36] Rosaria Conte and Cristiano Castelfranchi. *Cognitive and Social Action*. UCL Press, London, 1995.

[37] Antonio R. Damasio. The somatic marker hypothesis and the possible functions of the prefrontal cortex. *Philosophical Transactions of the Royal Society of London (series B)*, 351(1346):1413—1420, 1996.

[38] Charles R. Darwin. *The expression of emotions in man and animals*. Murray, London, 1872.

[39] D. Davidson. Actions, reasons, and causes (1963). In *Essay on Actions and Events*, essay 1, pages ?–? Oxford University Press, Oxford, 2nd edition, 2001.

[40] D. Davidson. *Essay on Actions and Events*. Oxford University Press, Oxford, 2nd edition, 2001.

[41] D. Davidson. Intending (1978). In *Essay on Actions and Events*, essay 5, pages ?–? Oxford University Press, Oxford, 2nd edition, 2001.

[42] F. de Rosis, C. Pelachaud, I. Poggi, V. Carofiglio, and B. De Carolis. From greta's mind to her face: modelling the dynamics of affective states in a conversational embodied agent. *International Journal of Human-Computer Studies*, 59:81–118, 2003.

[43] Daniel C. Dennett. *Darwin's Dangerous Idea: Evolution and the Meanins of Life*. Simon & Schuster, 1995.

[44] J. Drury, C. Cocking, and S. Reicher. Everyone for themselves? a comparative study of crowd solidarity among emergency survivors. *British Journal of Social Psychology*, 48:487–506, 2009.

[45] P. Ekman, R.W. Levenson, and W. Friesen. Autonomic nervous system activity distinguishes among emotions. *Science*, 221:1208–1210, 1983.

[46] Pascal Engel. Believing, holding true, and accepting. *Philosophical Explorations*, 1(2):140–151, 1998.

[47] R. L. Epstein. *The Semantic Foundations of Logic*, volume 1: *Propositional Logic*. Kluwer Academic Publishers, 1990.

[48] Ronald Fagin, Joseph Y. Halpern, Yoram Moses, and Moshe Y. Vardi. *Reasonning about Knowledge*. MIT Press, 1995.

[49] Luis Fariñas del Cerro, David Fauthoux, Olivier Gasquet, Andreas Herzig, Dominique Longin, and Fabio Massacci. LOTReC: a generic prover for modal and description logics. In R. Goré, A. Leitsch, and T. Nipkow, editors, *Int. Joint Conf. on Automated Reasonning (IJCAR'01), Siena, Italy, June 18–23*, number 2083 in LNAI, pages 453–458. Springer-Verlag, 2001.

[50] Luis Fariñas del Cerro, Andreas Herzig, Dominique Longin, and Omar Rifi. Belief reconstruction in cooperative dialogues. In Fausto Giunchiglia, editor, *Proc. 8th Int. Conf. on Artificial Intelligence: Methodology, Systems, and Applications (AIMSA'98), Sozopol, Bulgaria, September 21–23*, volume 1480 of *LNAI*, pages 254–266. Springer-Verlag, 1998.

[51] FIPA (Foundation for Intelligent Physical Agents). FIPA Communicative Act Library Specification. http://www.fipa.org/repository/aclspecs.html, 2002.

[52] N. Fornara, F. Vigan‘o, and M. Colombetti. Agent communication and artificial institutions. *Journal of Autonomous Agents and Multi-Agent Systems (JAAMAS)*, 14:121—-142, 2007.

[53] Nicoletta Fornara and Marco Colombetti. Operational Specification of a Commitment-Based Agent Communication Language. In Cristiano Castelfranchi and Lewis W. Johnson, editors, *Proceedings First Int. Joint Conf. on Autonomous Agents and MultiAgent Systems (AAMAS-2002)*, volume 2, pages 535–542. ACM Press, 2002.

[54] N. H. Frijda. *The Emotions*. Cambridge University Press, 1986.

[55] Olivier Gasquet, Andreas Herzig, Dominique Longin, and Mohamed Sahade. LoTREC: Logical Tableaux Research Engineering Companion. In Bernhard Beckert, editor, *International Conference on Automated Reasoning with Analytic Tableaux and Related Methods (TABLEAUX 2005), Koblenz, September 14–17, Germany*, pages 318–322. Springer Verlag, 2005.

[56] Benoit Gaudou, Andreas Herzig, and Dominique Longin. A Logical Framework for Grounding-based Dialogue Analysis. *Electronic Notes in Theoretical Computer Science (ENTCS)*, 157(4):117–137, 2006.

[57] Benoit Gaudou, Andreas Herzig, and Dominique Longin. A Logical Framework for Grounding-based Dialogue Analysis. *Proceedings of the Third IJCAI International Workshop on Logic and Communication in Multi-Agent Systems (LCMAS), Edinburgh, Scotland, UK, 01/08/2005*, van der Hoek, Wiebe and Lomuscio, Alessio and de Vink, Erik and Wooldridge, Mike (eds.), 2006.

[58] Benoit Gaudou, Andreas Herzig, and Dominique Longin. Grounding and the expression of belief. In Patrick Doherty, John Mylopoulos, and Christopher A. Welty, editors, *International Conference on Principles of Knowledge Representation and Reasoning (KR 2006), Windermere, UK, 02/06/2006-05/06/2006*, pages 211–229. AAAI Press, 2006.

[59] Benoit Gaudou, Andreas Herzig, and Dominique Longin. Logical formalization of social commitments: Application to agent communication languages (long version of AAMAS 2009). Rapport de recherche

IRIT/RR–2009-14–FR, IRIT, Université Paul Sabatier, Toulouse, mars 2009.

[60] Benoit Gaudou, Andreas Herzig, and Dominique Longin. Logical formalization of social commitments: Application to Agent Communication Languages (poster). In Keith Decker and Jaime Sichman, editors, *International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS), Budapest, Hungary, 10/05/2009-15/05/2009*, volume 2, pages 1293–1294. IFAAMAS, 2009.

[61] Benoit Gaudou, Andreas Herzig, and Dominique Longin. Group belief and grounding in conversation. In Alain Trognon, Martine Batt, Jean Caelen, and Denis Vernant, editors, *Logical Properties of Dialogue*, pages 59–96. Presses Universitaires de Nancy, http://www.univ-nancy2.fr/pun/, mars 2011.

[62] Benoit Gaudou, Andreas Herzig, Dominique Longin, and Matthias Nickles. A New Semantics for the FIPA Agent Communication Language based on Social Attitudes. In Gerhard Brewka, Silvia Coradeschi, Anna Perini, and Paolo Traverso, editors, *European Conference on Artificial Intelligence (ECAI 2006), Trento, Italy, 29/08/06-01/09/06*, pages 245–249. IOS Press, 2006.

[63] Benoit Gaudou, Dominique Longin, Emiliano Lorini, and Luca Tummolini. Anchoring Institutions in Agents' Attitudes: Towards a Logical Framework for Autonomous MAS. In *International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS), Estoril, Portugal, 12/05/08-16/05/08*, pages 728–735. ACM Press, 2008.

[64] M. P. Georgeff, B. Pell, M. Pollack, M. Tambe, and M. Wooldridge. The belief-desire-intention model of agency. In J. P. Muller, M. Singh, and A. S. Rao, editors, *Intelligent Agents V (LNAI)*. Springer Verlag, 1999.

[65] Edmund L. Gettier. Is justified true belief knowledge? *Analysis*, 23:121–123, 1963.

[66] R. Gordon. *The structure of emotions*. Cambridge University Press, New York, 1987.

[67] Jonathan Gratch and Stacy Marsella. Lessons from emotion psychology for the design of lifelike characters. *Journal of Applied Artificial Intelligence (special issue on Educational Agents - Beyond Virtual Tutors)*, 19(3-4):215–233, 2005.

[68] Nadine Guiraud, Dominique Longin, Emiliano Lorini, Sylvie Pesty, and Jérémy Rivière. The face of emotions: a logical formalization of expressive speech acts (regular paper). In *International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS), Taipei, Taiwan, 02/05/2011-06/05/2011*, pages 1031–1038, http://www.acm.org/, mai 2011. ACM Press.

[69] P. Hakli. Group beliefs and the distinction between belief and acceptance. *Cognitive Systems Research*, 7:286–297, 2006.

[70] D. Harel, D. Kozen, and J. Tiuryn. *Dynamic Logic*. MIT Press, Cambridge, 2000.

[71] D. Harel, D. Kozen, and J. Tiuryn. *Dynamic Logic*. MIT Press, Cambridge, 2000.

[72] S. Hareli and A. Rafaeli. Emotion cycles: On the social influence of emotion in organizations. *Research in Organizational Behavior*, 28:35–59, 2008.

[73] E. Hatfield, J.T. Cacioppo, and R.L. Rapson. *Emotional contagion*. Cambridge University Press, New-York, 1994.

[74] Raymond Reiter Hector J. Levesque, Yves Lespérance, Fangzhen Lin, and Richard B. Scherl. Golog: A logic programming language for dynamic domains. *Journal of Logic Programming*, 31:59–84, 1997.

[75] Andreas Herzig, Jérôme Lang, and Dominique Longin. I thought you didn't know! On belief revision in dynamic doxastic logic. Fifth int. conf. of Logic and Foundations of Game and Decision Theory (LOFT 5), Torino, Italy, June 28–30, 2002.

[76] Andreas Herzig, Jérôme Lang, Dominique Longin, and Thomas Polacsek. A logic for planning under partial observability. In Henry Kautz

and Bruce Porter, editors, *Proc. Seventeenth National Conf. on Artificial Intelligence (AAAI-00), Austin, Texas, Aug. 30–Sep. 3*, pages 768–773. AAAI Press, 2000.

[77] Andreas Herzig and Dominique Longin. Belief dynamics in cooperative dialogues. Third Int. Workshop on the Semantics and Pragmatics of Dialogue (Amstelogue'99), Amsterdam, Holland, May 7–9, 1999.

[78] Andreas Herzig and Dominique Longin. Belief dynamics in cooperative dialogues. *Journal of Semantics*, 17(2):91–118, 2000.

[79] Andreas Herzig and Dominique Longin. A topic-based framework for rational interaction. 7e conf. annuelle sur le Traitement Automatique des Langues Naturelles (TALN-2000), Lausanne, Suisse, Oct. 16–18, p. 431–434, 2000.

[80] Andreas Herzig and Dominique Longin. A logic of intention with cooperation principles and with assertive speech acts as communication primitives. In Cristiano Castelfranchi and W. Lewis Johnson, editors, *Proc. of the first Int. Joint Conf. on Autonomous Agent and Multi-Agent System (AAMAS 2002), Bologna, Italy, July 15–19*, volume 2, pages 920–927. ACM Press, 2002.

[81] Andreas Herzig and Dominique Longin. Sensing and revision in a modal logic of belief and action. In F. van Harmelen, editor, *Proc. of 15th European Conf. on Artificial Intelligence (ECAI 2002), Lyon, France, July 23–26*, pages 307–311. IOS Press, 2002.

[82] Andreas Herzig and Dominique Longin. Sensing and revision in a modal logic of belief and action (preliminary report). Technical Report 2002-01-R, Institut de Recherche en Informatique de Toulouse, `www.{InstitutdeRechercheenInformatiquedeToulouse}.fr/~Dominique.Longin/`, January 2002. 16 pages.

[83] Andreas Herzig and Dominique Longin. C&L intention revisited. In Didier Dubois, Chris Welty, and Mary-Anne Williams, editors, *Proc. 9th Int. Conf. on Principles of Knowledge Representation and Reasoning (KR 2004), Whistler, Canada, June 2–5*, pages 527–535. AAAI Press, 2004.

[84] Andreas Herzig, Dominique Longin, and Jacques Virbel. Towards an analysis of dialogue acts and indirect speech acts in a BDI framework. Fourth Int. Workshop on the Semantics and Pragmatics of Dialogue (Götalog-2000), Göteborg, Sueden, Feb. 1–2, 2000.

[85] Andreas Herzig, Emiliano Lorini, Jomi F. Hübner, Jonathan Ben-Naim, Olivier Boissier, Cristiano Castelfranchi, Robert Demolombe, Dominique Longin, Laurent Perrussel, and Laurent Vercouter. Prolegomena for a logic of trust and reputation. In Guido Boella, Gabriella Pigozzi, Munindar Singh, and Harko Verhagen, editors, *International Workshop on Normative Multiagent Systems (NorMAS), Luxembourg, 15/07/2008-16/07/2008*, pages 143–157. University of Luxembourg Press, 2008. ISBN: 2919940481.

[86] Andreas Herzig, Emiliano Lorini, Jomi F. Hübner, and Laurent Vercouter. A logic of trust and reputation. *Logic Journal of the IGPL, Normative Multiagent Systems*, 18(1):214–244, février 2010.

[87] J. Hintikka. *Knowledge and Belief: An Introduction to the Logic of the Two Notions*. Cornell University Press, Ithaca, 1962.

[88] W. James. What is an emotion? *Mind*, 9:188–205, 1884.

[89] A. Jones and M. J. Sergot. A formal characterization of institutionalised power. *Journal of the IGPL*, 4:429–445, 1996.

[90] H. Katsuno and A. Mendelzon. Propositional knowledge base revision and minimal change. *Artificial Intelligence*, 52(3):263–294, 1991.

[91] Hirofumi Katsuno and Alberto O. Mendelzon. On the difference between updating a knowledge base and revising it. In Peter Gärdenfors, editor, *Belief Revision*. Cambridge University Press, 1992.

[92] E. Lagerspetz. *The opposite mirrors*. Kluwer, Dordrecht, 1995.

[93] R. Lane and L. Nadel, editors. *The cognitive neuroscience of emotions*. Oxford University Press, New York, 2000.

[94] Richard Lazarus. Thoughts on the Relation between Emotion and Cognition. *American Psychologist*, 37(9):1019–1024, 1982.

[95] Richard Lazarus. The Cognition–Emotion Debate: a Bit of History. In Tim Dalgleish and Mick Power, editors, *Handbook of Cognition and Emotion*, pages 3–20. John Wiley & Sons, New York, 1999.

[96] Richard S. Lazarus. *Emotion and Adaptation*. Oxford University Press, 1991.

[97] Van Minh Le, Carole Adam, Richard Canal, Benoit Gaudou, Tuong Vinh Ho, and Patrick Taillander. Simulation of the emotion dynamics in a group of agents in an evacuation situation. In *Principles and Practice of Multi-Agent Systems - 13th International Conference, PRIMA 2010, Kolkata, India, November 12–15, 2010*, 2010.

[98] C. Lebiere and J. R. Anderson. A connectionist implementation of the act-r production system. In *Fifteenth Annual Conference of the Cognitive Science Society*, pages 635–640, 1993.

[99] G. Loewenstein. Emotions in economic theory and economic behavior. *American Economic Review*, 90(2):426–432, 2000.

[100] Dominique Longin. *Interaction rationnelle et évolution des croyances dans le dialogue : une logique basée sur la notion de topique*. PhD thesis, Université Paul Sabatier, Toulouse, France, November 1999. `www.{InstitutdeRechercheenInformatiquedeToulouse}.fr/~Dominique.Longin/`.

[101] Dominique Longin. Évolution des croyances au sein d'une théorie de l'intentionnalité : application au dialogue coopératif orienté tâches. Rencontres des Jeunes Chercheurs en Intelligence Artificielle (RJCIA-2000), Lyon, France, 10–13 sept., 2000.

[102] Dominique Longin. Des raisons qu'ont certains actes à être indirects. *Psychologie de l'Interaction*, 21–22:237–258, 2006.

[103] Dominique Longin and Éric Raufaste. Actes de langage indirects : co-construction d'un modèle logico-psychologique. In A. Herzig, B. Chaib-Draa, and Ph. Mathieu, editors, *Proc. 2ndes Journées Francophones Modèles Formels de l'Interaction (MFI'03), Lille, France, 20–22 mai*, pages 169–178. Cépaduès-Éditions, 2003.

[104] Dominique Longin and David Sadek. Dialogue et dynamique des croyances. In Henri Prade, Robert Jeansoulin, and Catherine Garbay, editors, *Le temps, l'espace et l'évolutif en Sciences du Traitement de l'Information*, pages 345–359. Cépaduès-Éditions, `http://www.cepadues.com/`, 2000.

[105] E. Lorini, D. Longin, and E. Mayor. A logical analysis of responsibility attribution: emotions, individuals and collectives. *Journal of Logic and Computation*, 24(6):1313–1339, December 2014.

[106] E. Lorini and F. Schwarzentruber. A logic for reasoning about counterfactual emotions. *Artificial Intelligence*, 175:814–847, 2011.

[107] Emiliano Lorini and Dominique Longin. A logical account of institutions: from acceptances to norms via legislators. In G. Brewka and J. Lang, editors, *International Conference on Principles of Knowledge Representation and Reasoning (KR), Sidney, Australia, 16/09/2008-19/09/2008*, pages 38–48. AAAI Press, 2008.

[108] Emiliano Lorini, Dominique Longin, and Benoit Gaudou. The institutional dimension of speech acts: a logical approach based on the concept of acceptance. Research report IRIT/RR–2008-9–FR, IRIT, Paul Sabatier University, Toulouse, FRANCE, February 2008. `ftp://ftp.irit.fr/IRIT/LILAC/2008_lorini_et_al_RR--2008-9--FR.pdf`.

[109] Emiliano Lorini, Dominique Longin, and Benoit Gaudou. Anchoring the institutional dimension of speech acts in agents' attitudes: a logical approach (regular (long) paper). In Tru Cao, Ralf-Detlef Kutsche, and Akim Demaille, editors, *IEEE International Conference on Research, Innovation and Vision for the Futur (RIVF), Da Nang City, Viet Nam, 13/07/2009-17/07/2009*, pages 65–72. IEEE, 2009.

[110] Emiliano Lorini, Dominique Longin, Benoit Gaudou, and Andreas Herzig. The logic of acceptance: grounding institutions on agents' attitudes. *Journal of Logic and Computation*, 19(6):901–940, 2009.

[111] D. Makinson. On the formal representation of rights relations. *Journal of Philosophical Logic*, 15(4):403–425, 1986.

[112] J. J. Meyer. A different approach to deontic logic: Deontic logic viewed as a variant of dynamic logic. *Notre Dame Journal of Formal Logic*, 29(1):109–136, 1988.

[113] John Jules Meyer. Reasoning about emotional agents. In R. López de Mántaras and L. Saitta, editors, *16th European Conf. on Artif. Intell. (ECAI)*, pages 129–133, 2004.

[114] John-Jules Meyer, Frank de Boer, Rogier van Eijk, Koen Hindriks, and Wiebe van der Hoek. On programming karo agents. *Logic Journal of the IGPL*, 9(22):261–272, 2001.

[115] Andrew Ortony, G.L. Clore, and A. Collins. *The cognitive structure of emotions*. Cambridge University Press, Cambridge, MA, 1988.

[116] Catherine Pelachaud. Modelling multimodal expression of emotion in a virtual agent. *Philosophical transactions of the Royal society B*, 364:3539–3548, 2009.

[117] J. Piaget. Les émotions. In B. Rimé and K. Scherer, editors, *Les relations entre l'intelligence et l'affectivité dans le developpement de l'enfant*, pages 75–95. Delachaux et Niestlé, Neuchâtel-Paris, 1989.

[118] David Pitt. Mental Representation. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy (Fall 2013 Edition)*. `http://plato.stanford.edu/archives/fall2013/entries/mental-representation/`, 2013.

[119] Sally Popkorn. *First Steps in Modal Logic*. Cambridge University Press, 1994.

[120] E.L. Quarantelli. The sociology of panic. In Smelser and Baltes, editors, *International Encyclopedia of the Social and Behavioural Sciences*, pages 11020–11023. Pergamon Press, New York, 2001.

[121] Anand S. Rao. Agentspeak(l): Bdi agents speak out in a logical computable language. In *Proceedings of the 7th European Workshop on Modelling Autonomous Agents in a Multi-agent World : Agents Breaking Away: Agents Breaking Away*, MAAMAW '96, pages 42–55, Secaucus, NJ, USA, 1996. ACM Press.

[122] Anand S. Rao and Michael P. Georgeff. Modeling rational agents within a BDI-architecture. In J. A. Allen, R. Fikes, and E. Sandewall, editors, *Proceedings Second Int. Conf. on Principles of Knowledge Representation and Reasoning (KR'91)*, pages 473–484. Morgan Kaufmann Publishers, 1991.

[123] Anand S. Rao and Michael P. Georgeff. An abstract architecture for rational agents. In Bernhard Nebel, Charles Rich, and William Swartout, editors, *Proceedings Third Int. Conf. on Principles of Knowledge Representation and Reasoning (KR'92)*, pages 439–449. Morgan Kaufmann Publishers, 1992.

[124] Anand S. Rao and Michael P. Georgeff. Decision procedures for BDI logics. *Journal of Logic and Computation*, 8(3):293–343, 1998. Special Issue: Computational and Logical Aspects of Multiagent Systems, Oxford univerty Press.

[125] E. Raufaste. *Les mécanismes cognitifs du diagnostic médical: optimisation et expertise*. Presses Universitaires de France, Paris, 2001.

[126] Ray Reiter. A logic for default reasoning. *Artificial Intelligence Journal*, 13:81–132, 1980.

[127] Howard Robinson. Dualism. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy (Winter 2012 Edition)*. `http://plato.stanford.edu/archives/win2012/entries/dualism/`, 2012.

[128] J. Sabater and C. Sierra. Review on Computational Trust and Reputation Models. *Artificial Intelligence*, 24:33–60, 2005.

[129] M. D. Sadek. *Attitudes mentales et interaction rationnelle : vers une théorie formelle de la communication*. PhD thesis, Université de Rennes I, Rennes, France, 1991.

[130] M.D. Sadek. A study in the logic of intention. In Bernhard Nebel, Charles Rich, and William Swartout, editors, *Proceedings Third Int. Conf. on Principles of Knowledge Representation and Reasoning (KR'92)*, pages 462–473. Morgan Kaufmann Publishers, 1992.

[131] M.D. Sadek. Dialogue acts are rational plans. In M.M. Taylor, F. Néel, and D.G. Bouwhuis, editors, *The structure of multimodal dialogue*,

pages 167–188. John Benjamins publishing company, 2000. From ESCA/ETRW, Workshop on The Structure of Multimodal Dialogue (Venaco II), 1991.

[132] K. R. Scherer, A. Schorr, and T. Johnstone, editors. *Appraisal Processes in Emotion : Theory, Methods, Research*. Oxford University Press, New York, 2001.

[133] J. R. Searle. *Speech acts: An essay in the philosophy of language*. Cambridge University Press, New York, 1969.

[134] J. R. Searle. Is the brain a digital computer? *Proceedings and Addresses of the American Philosophical Association*, 64, 1990.

[135] J. R. Searle. *The Construction of Social Reality*. The Free Press, New York, 1995.

[136] John R. Searle. *Expression and Meaning. Studies on the Theory of Speech Acts*. Cambridge University Press, 1979.

[137] John R. Searle. *Intentionality: An essay in the philosophy of mind*. Cambridge University Press, 1983.

[138] John R. Searle. *Minds, Brains and Science*. British Broadcasting Corporation, 1984.

[139] John R. Searle. *Du cerveau au savoir*. Hermann, 1985.

[140] S. Shapiro, Yves Lespérance, and Hector J. Levesque. Specifying communicative multi-agent systems with ConGolog. In *Working notes of the AAAI fall symposium on Communicative Action in Humans and Machines*, pages 75–82. AAAI Press, 1997.

[141] M. P. Singh. An ontology for commitments in multiagent systems. *Artificial Intelligence and Law*, 7:97–113, 1999.

[142] Munindar P. Singh. Agent Communication Languages: Rethinking the Principles. *Computer*, 31(12):40–47, December 1998.

[143] Munindar P. Singh. A Social Semantics for Agent Communication Languages. In Frank Dignum and Mark Greaves, editors, *Issues in Agent Communication, IJCAI'99 Workshop on "Agent Communication Languages"*, number 1916 in LNAI, pages 75–88. Springer-Verlag, 2000.

[144] Robert C. Solomon. The Philosophy of Emotion. In Michael Lewis, Jeannette M. Haviland-Jones, and Lisa Feldman Barrett, editors, *Handbook of Emotions*, pages 3–16. The Guilford Press, New York, 3rd edition, 2008.

[145] R. Stalnaker. *Inquiry*. MIT Press, Cambridge, 1984.

[146] S.S. Tomkins. Affect theory. In K.R. Scherer and P. Ekman, editors, *Approaches to emotion*, pages 163–196. Erlbaum, Hillsdale, NJ, 1984.

[147] Alain Trognon, Christine Sorana, Martine Batt, and Dominique Longin. Comment identifie-t-on un théorème-en-acte en logique interlocutoire ? In Maryvonne Merri, editor, *Activité humaine et conceptualisation – Questions à Gérard Vergnaud*, pages 763–782. Presses Universitaire du Mirail, 2007.

[148] Alain Trognon, Christine Sorsana, Martine Batt, and Dominique Longin. Peer interaction and problem solving: one example of a logical-discursive analysis of a process of joint decision making. *The European Journal of Developmental Psychology*, 5(5):623–643, 2008.

[149] Jason Tsai, Natalie Fridman, Emma Bowring, Matthew Brown, Shira Epstein, Gal Kaminka, Stacy Marsella, Andrew Ogden, Inbal Rika, Ankur Sheel, Matthew Taylor, Xuezhi Wang, Avishay Zilka, and Milind Tambe. ESCAPE - Evacuation Simulation with Children, Authorities, Parents, Emotions, and Social comparison. In Tumer, Yolum, Sonnenberg, and Stone, editors, *Proc. of 10th Int. Conf. on Autonomous Agents and Multiagent Systems – Innovative Applications Track (AAMAS 2011), May, 2–6, 2011, Taipai, Taiwan*, 2011.

[150] R. Tuomela. *The Philosophy of Social Practices: A Collective Acceptance View*. Cambridge University Press, Cambridge, 2002.

[151] Raimo Tuomela. Group beliefs. *Synthese*, 91:285–318, 1992.

[152] Raimo Tuomela. Belief versus Acceptance. *Philosophical Explorations*, 2:122–137, 2000.

[153] P. Turrini, J.-J. Ch. Meyer, and C. Castelfranchi. Coping with shame and sense of guilt: a dynamic logic account. *Journal of Autonomous Agents and Multi-Agent Systems*, 20(3), 2010.

[154] W. van der Hoek, W. Jamroga, and M. Wooldridge. Towards a theory of intention revision. *Synthese*, 155(2):265–290, 2007.

[155] H. van Ditmarsch, W. van der Hoek, and B. Kooi. *Dynamic Epistemic Logic*. Kluwer Academic Publishers, 2007.

[156] Daniel Vanderveken. *Principles of language use*, volume 1 of *Meaning and Speech Acts*. Cambridge University Press, 1990.

[157] Mario Verdicchio and Marco Colombetti. A Logical Model of Social Commitment for Agent Communication. In *Proceedings Second Int. Joint Conf. on Autonomous Agents and MultiAgent Systems (AAMAS-2003)*, pages 528–535. ACM Press, 2003.

[158] Jacques Virbel. Contributions de la théorie des actes de langage à une taxinomie des consignes. In J. Virbel, J-M. Cellier, and J-L. Nespoulous, editors, *Cognition, Discours procédural, Action*, volume II. PRESCOT, 1999.

[159] D. N. Walton and E. C. Krabbe. *Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning*. State University of New-York Press, NY, 1995.

[160] Michael Wooldridge. *Reasoning about Rational Agents*. MIT Press, 2000.

[161] R.B. Zajonc. Feeling and Thinking – Preferences Need No Inferences. *American psychologist*, 35(2):151–175, February 1980.

[162] Robert B. Zajonc. Feeling and Thinking; closing the Debate over the Independence of Affect. In J.P. Forgas, editor, *Feeling and Thinking: the Role of Affect in Social Cognition*, pages 31–59. Cambridge University Press, 2000.

[163] Marcel Zeelenberg, W. W. van Dijk, and A. S. R. Manstead. Reconsidering the relation between regret and responsibility. *Organizational Behavior and Human Decision Processes*, 74:254–272, 1998.