THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par *l'Université Toulouse III - Paul Sabatier*
Discipline ou spécialité : *Informatique - Intelligence Artificielle*

Présentée et soutenue par *Gaudou Benoit*
Le *10 juillet 2008*

Titre : *Formalizing social attitudes in modal logic*

JURY
Leon van der Torre - Professeur, Université du Luxembourg - Rapporteur
Denis Vernant - Professeur, UPMF Grenoble - Rapporteur
Andreas Herzig - Directeur de Recherche CNRS, IRIT - Directeur de thèse
Dominique Longin - Chargé de Recherche CNRS, IRIT - Directeur de thèse
Marie-Pierre Gleizes - Professeur, UPS - Présidente du jury
Hans van Ditmarsch - Maître de conférence, Université d'Otago, Nouvelle-Zélande - Membre
Laurent Vercouter - Maître de conférence - Ecole des Mines de Saint-Etienne - Membre

---

Ecole doctorale : *Mathématiques, Informatique, et Télécommunications de Toulouse*
Unité de recherche : *Institut de Recherche en Informatique de Toulouse*
Directeur(s) de Thèse : *Andreas Herzig - Dominique Longin*
Rapporteurs : *Leon van der Torre - Denis Vernant*

Formalizing social attitudes in modal logic

Benoit Gaudou

Under the direction of Andreas Herzig and Dominique Longin

Reviewers:
Leon van der Torre
Denis Vernant

Examiners:
Nicholas Asher
Hans van Ditmarsch
Marie-Pierre Gleizes
Laurent Vercouter

Contents

Remerciements

Tout d'abord, je voudrais dire un très grand merci à Léon van der Torre et Denis Vernant pour avoir accepté d'être rapporteurs de ma thèse, et particulièrement pour la qualité de leur rapport malgré le temps très court qui leur avait été imparti. Un grand merci également pour les questions et les discussions qui ont suivi et qui m'ont apporté une vision plus large et plus profonde sur ce vaste domaine d'étude.

Je remercie Marie-Pierre Gleizes, Laurent Vercouter et Hans van Ditmarch d'avoir accepté d'être membre du jury, malgré les différences de point de vue sur les Systèmes Multi-Agents. Je suis très heureux qu'un jury aussi éclectique ait assisté à ma soutenance. Je remercie également Nicholas Asher d'avoir accepté d'être membre et je regrette qu'il n'ait pu être présent lors de la soutenance, je suis sûr que ses questions et remarques auraient été passionnantes.

Je remercie particulièrement mes directeurs de thèse avec qui j'ai eu énormément de plaisir à travailler pendant ces quatre années. Merci à Andreas Herzig de m'avoir accepté comme thésard alors que je débarquais de nulle part après mon stage de DEA de planification de mouvement probabiliste appliquée aux molécules et de m'avoir remis sur le droit chemin en m'introduisant à la logique modale par la lecture de la bible selon Chellas. Un grand merci également à Dominique Longin pour m'avoir inculqué un peu de son savoir infini sur le Latex, pour m'avoir initié aux joies des actes de langage et du thé chinois et pour m'avoir appris comment réduire de moitié un article sans toucher à son contenu la nuit précédant la deadline pour le finir juste 5 minutes avant la cloture des soumissions.

Je remercie les personnes avec qui j'ai pu travailler de près, notamment Matthias Nickles et Luca Tummolini avec qui j'ai eu le plaisir de collaborer, ou de plus loin, à savoir les membres des équipes LILaC (et en particulier à la myriade de thésards d'Andi qui se sont dispersé ensuite aux quatre coins du monde) et RPDMP de l'IRIT. Des remerciements particulièrement appuyés à Carole Adam et Emiliano Lorini qui m'ont initié à des passionnants sujets. De futurs travaux en commun pourraient porter sur des problématiques aussi farfelues (quoique ...) que : Comment donner des émotions à une institution ? Ou réciproquement : Comment institutionaliser ses émotions ?

Comment ne pas remercier ceux et celles qui m'ont permis de me détendre entre deux articles et de ma changer les idées de mes recherches, en me donnant l'occasion d'enseigner. Un grand merci à tous les enseignants, moniteurs et

ATERs de l'ENSEEIHT ainsi qu'aux étudiants qui ont dû me supporter, ce fut pour moi une expérience humaine particulièrement enrichissante.

Et comme le travail d'un doctorant n'est pas celui d'un fonctionnaire (on n'a pas la sécurité de l'emploi ...) travaillant derrière son bureau (... ah si ça on le fait ...) 5 jours par semaine à heures fixes, je tiens à remercier tous ceux sur qui mon travail a pu resurgir dans la vie de tous les jours (et qui ont donc d'une manière ou d'une autre influencé mon travail), à savoir mes proches, ma famille et mes amis. Je ne cite pas de noms histoire de n'oublier personnes. En particulier, mille mercis à celle qui a supporté les : "Ce soir/Cette nuit/Ce week-end/Ce jour férié/Ces vacances (rayer la mention inutile), je ne peux pas, j'ai du boulot, désolé ...".

Chapter 1

Introduction

The study of the human beings' mind is certainty one of the oldest and most important objects of study in philosophy, but also one of the most polemic. We can distinguish two main starting points to study mind. First we can study mind by introspection. Every human being has at his disposal a mind and his own beliefs, desires, *etc.*... He is aware of his own mind but he has no means to access the others' minds: nobody can verify that somebody else really has a mind and that this mind has the same structure and the same mental states as his, or that these mental states have the same properties. This leads to the solipsism problem (since Gorgias in lost works or (Descartes, 1968)) that can be summarized by: my mind is the only thing of which I know it exists.

A way to get around this problem and to analyse another human being's mind is to *ascribe* mental attitudes to him, depending on his behavior. Of course, a correct analysis requires exact knowledge about environmental stimuli to which he has reacted (which is a really hard problem). Moreover his behavior is biased by the observer, in the sense that he can behave in some way with an observer and in another way with another observer.

Multi-Agent Systems (MASs) are a uniform way to represent human beings or other entities having some autonomy and being able to interact with the environment and in particular with other agents. Such systems gained increasing importance during the last years. It turned out to be fruitful to analyze both human and artificial agents in MASs in terms of mental attitudes. When applied to MASs, above remarks compel us to examine agents' mental attitudes from both points of view: the private and the public one. In particular they highlight the need to identify the group of agents observing the mental attitude under concern. If a given group of agents is aware of some propositions (representing some observations) while others are not, we can similarly to individual agents utter that this group has a belief that other groups do not have.

In most (if not all) cases, group belief results from the agents' interaction. Consider the example of three rational agents in a company. Agent 0 thinks privately for some reasons that his boss (agent 2) is smart. But this idea is not widespread in his department: agent 0 meets agent 1, a very charismatic

agent who often claims publicly that his boss (agent 2) is dumb. They discuss about their boss and agent 0 asserts that he is really a moron (for some social reasons) and of course agent 1 confirms. At this moment the boss comes and enters the conversation. Soon they get to discuss about himself, and agent 0 congratulates himself by asserting he is smart. Agent 2 and agent 1 (given the boss' attendance) express their agreement. These three agents together seem to have reached a stable state about their boss. The first aim of this dissertation is to define a logic of group belief that is able to represent such situations. Note that the notion of common belief that is widely studied in theoretical computer science and artificial intelligence is not appropriate for our purposes. Indeed, if some group has a common belief then every subgroup of that group also holds that common belief: if the group made up of 0, 1 and 2 commonly believes that 2 is smart then the subgroup made up of 0 and 1 also commonly believes so.

As in our example, group belief typically results from a dialogue between two or more agents. It is important to note that in our example agent 0 expresses contradictory points of view, depending on his hearers, and possibly distinct from his private beliefs. As we are interested by what is public in a dialogue, we have to make precise which group of agents makes up the public: as illustrated above, a speaker's behavior depends on who can hear what he says. Thus the second aim of this dissertation is to present a formal account of dialogues with different groups of hearers. We argue that group belief is a useful tool in the analysis of dialogues between agents. We will focus on Agent Communication Languages (ACLs).

Indeed the dichotomy presented above will be present all along the thesis. We will always make the distinction between public and private layer, in particular between between agents' individual beliefs(that are private) and group belief (that is public). But this group belief is the belief of the given group and thus remains private outside of this group, in particular for other groups. We will give a logical formalization of group belief in modal logic. Concerning the study and formalization of mental attitudes, obligations, time or actions, modal logic is a widespread tool for purely theoretical studies (such as in philosophy) but also for practical-oriented studies, (such as in artificial intelligence or in MASs).

The main contribution of the dissertation is the formalization in modal logic of the group belief in a non-reductionist sense and its integration in a logical BDI framework (and in particular the comparison with common belief as usually defined). Note that as far as we are aware, there does not exist another logical formalization of the group belief in this sense with a standard possible worlds semantics. The second contribution is the extension of this logic to a logic of the group acceptance by introducing an institutional context. It allows to formalize informal institutions anchored in agents attitudes. The last contribution is the application of the group belief logic to formalize Agent Communication Languages. We take as starting hypothesis that by performing a speech act, an agent publicizes some mental attitudes. We argue that we can use group belief operator to give a new kind of semantics to speech acts, bridging the gap between mentalist and social approaches of ACLs. We thus give a unification

of FIPA mentalist ACL and Walton & Krabbe's and Colombetti et al.'s social approaches.

At this point, it is interesting to discuss our choice of formalism, that is modal logic. As mentioned above, we aim at formalize group belief and its links with individual mental attitudes and in particular its (non-)link with individual belief. Modal logic is the classical logical formalism to represent these notions. Moreover this formalism is often used in the definition of speech acts semantics (for example in FIPA-ACL). Given our aims, modal logic appears the most natural and powerful approach to represent studied social notions and highlight formal links with individual attitudes. Nevertheless we are aware this approach is really hard to implement and thus that logics are not used directly practically. Moreover we admit that our language lacks expressivity comparing other formalisms. In particular we do not have any notion of degrees in our operators, they only are binary (fully true or fully false), which could be a big limitations in implemented systems considering social notions such as trust or reputation.

We begin this dissertation by introducing in Chapter 2 the issue of collective Intentionality and in particular of group belief and its link with group acceptance. We will make a clear distinction between the reductionist view of group belief (in the sense of common belief, *i.e.* with a group belief depending of the individual beliefs) and the proper group belief (in Gilbert's sense (Gilbert, 1987), *i.e.* with a group belief independent from individual beliefs). We summarize the debate around Gilbert's group belief to clarify its nature: is it a particular kind of belief or rather a kind of acceptance? This discussion and the above one about reductionist and non-reductionist approaches allow to highlight main features of group belief in view of formalization.

Chapter 3 is dedicated to the logical formalization of proper group belief in a modal logic. We highlight its properties and link it with other mental attitudes such as individual belief, choice and intention, and with actions. We also compare it with the common belief operator and point out main differences between these two views of the belief of a group of agents. The material in this chapter is based on a chapter that was contributed to the book "Language, Cognition, Interaction" (Gaudou, Herzig, and Longin, 2008).

We extend our logical framework to take into account group acceptance in Chapter 4. We are interested in the acceptance of a proposition by a group as followers of an institution. We show that this allows to anchor the institutional rules in public mental attitudes of groups of agents. The main idea is that an institution rests on the acceptance of agents following this institution. We explore informal institutions but give some starting ideas to take into account formal institutions, too. The material of this chapter was published at the 2008 International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS'2008) (Gaudou et al., 2008).

We then present in Chapter 5 an overview of the state of the art in theories of dialogue. We briefly survey Speech Act Theory, and focus on its application to artificial agents with ACLs. We present two opposed approaches: the mentalist and the social approaches. Once again the above dichotomy appears:

a distinction exists between mentalist approaches, based on agents' mental attitudes, and social approaches that are based on what is public (in particular commitments). We propose a new one to bridge the gap between these two approaches, attempting to bring together advantages of both. We thus study the grounding process and link it with group belief.

In Chapter 6 we present how to formalize the standard mentalist ACL (FIPA-ACL (FIPA, 2002a)) with our framework. In particular we show that we can catch interaction protocols used to manage dialogues such as the Contract Net Protocol. This chapter has been published for a large part in a paper at the 2006 International Conference on Knowledge Representation and Reasoning (KR'2006) (Gaudou, Herzig, and Longin, 2006a) and a paper at the 2006 European Conference on Artificial Intelligence (ECAI'2006) (Gaudou et al., 2006).

Finally, Chapter 7 contains a formalization of the other classical approach to dialogue, that is the social approach based on commitments. We thus propose a representation of commitments in our framework. Part of this chapter has been published in (Gaudou, Herzig, and Longin, 2006b).

Chapter 2

Group belief: a state of the art

2.1 Introduction

Individual belief and other individual intentional attitudes have been deeply studied by philosophers and logicians. Since the Greeks knowledge and its link with belief has been the subject of interest of philosophers such as Plato (Platon, 1999): for example, knowledge was viewed as a justified true belief. The Middle Ages were also the time of great interest for epistemic logic around various problems such as the formulation of the epistemic conception of entailment-propositions, inferences with epistemic and doxastic formula or the links between concepts of truth, faith, knowledge and belief (Gochet and Gribomont, 2006). More details about epistemic logic in this period can be found in (Boh, 1993). In the twentieth century, the development of mathematical/logical formalisms such as the possible world semantics for modal logic (Kripke, 1963) triggered a revival of interest for the study of those concepts and a proliferation of formalization of various mental mental states. The main attitudes that were investigated are knowledge and belief with (Hintikka, 1962) as forerunner, followed by, among a lot of others, (Fagin et al., 1995), (Lenzen, 1980) and (Meyer and van der Hoek, 1995), or Intention (Cohen and Levesque, 1990a). We will describe in details in the following chapter a classical formalization of belief.

Although individual belief and shared belief, *i.e.* belief shared by every member of a group of agents (Tuomela, 2004; Schelling, 1960; Scheff, 1967; Lewis, 1969; Schiffer, 1972) were investigated for decades, collective intentionality (with (Searle, 1995)), *i.e.* intentionality ascribed to a group of agents, in general and group belief (with (Gilbert, 1987)) in particular has become an active field only recently. In economics, the related question of Social Choice has been introduced by Arrow in (Arrow, 1951) in order to merge both welfare economics and voting theory. In this thesis, the focus of investigation is collective belief. We will nevertheless address other doxastic attitudes such as acceptance

to better understand and characterize belief.

In natural language beliefs are commonly ascribed to groups as in the following example.

EXAMPLE. *[(Tuomela, 1992)] The government believes that war against Iraq will begin soon.*

But this attribution is not so obvious and one has to justify its relevance. We will argue for it in Section 2.2. This will give us the opportunity to introduce some basic important concepts such as the ones of intentional states, doxastic attitude in general and belief in particular.

Some preliminary investigations to describe collective belief have been led by Anthony Quinton (Quinton, 1976) and Emile Durkheim (Durkheim, 1982) among others. But the first study and characterization of group belief as a whole is due to Margaret Gilbert (Gilbert, 1987; Gilbert, 1989) and described and discussed in Section 2.3. Her account will serve as starting point and reference all along the thesis. This account has been the object of a lot of comments and criticisms, in particular by whom she named *Rejectionnists*, *e.g.* (Meijers, 1999), (Wray, 2001) and (Tuomela, 2000). Criticisms concern the following point: the phenomenon described by Gilbert as a kind of belief is rather a kind of acceptance. This dispute, presented in Section 2.4.3, will lead to a deep and interesting analysis of both notions. Moreover it will help us to identify main features of group belief (Section 2.5).

2.2 Important preliminary notions

For a long time Intentionality was exclusively ascribed to individual agents. For example, for (Goldman, 1987) (cited from (Tollefsen, 2002)):

> "Knowers are individuals, and knowledge is generated by mental processes and lodged in the mind-brain."

In this section we define and defend the idea of a collective Intentionality and in particular that collective belief has to be define in a non-reductionist way.

2.2.1 Intentionality

For Searle, following Brentano (Brentano, 1995) and Husserl, "Intentionality is that property of many mental states and events by which they are directed at or about or of objects and states of affairs in the world" (Searle, 1983, p. 1). This characterization can be applied to many mental states such as belief, intention and some emotions (joy, hope...): we believe that the earth is flat, we have the intention to go to the dentist or we feel joy to meet a friend. It would be problematic to utter: "I believe and I intend.". Of course hearers would ask: "But what do you believe? What do you intend to do?". It should always be possible to exhibit the object of an intentional attitude, what this attitude is about.

The concept of Intentionality has been used as key criterion by Brentano to distinguish mental (which are Intentional) from physical facts (which are not). Searle among others questions this distinguishing criterium: Intentionality can be applied to many mental states such as belief, intention or some emotions (joy, hope...). But not every mental attitude can be qualified to be Intentional. For example, we can be in a bad mood today or we can be in an intense exaltation state without being able to name the object of these feelings. Moreover it can be argued that some physical objects could be called Intentional. For example, a picture is always a picture of someone or something and thus by extending Searle's characterization we can say that it is also intentional. Thus Intentionality cannot be the key distinguishing criterium between physical and mental objects. In the sequel the word Intentionality will nevertheless always describe mental states. Moreover Intentionality should not be confused with awareness: some conscious mental attitudes can be unintentional (a conscious exaltation for example) and we are not aware of all our intentional mental attitudes (we have a lot of beliefs and desires that stay unconscious).

It is also highly important for this thesis to highlight the distinction between Intentionality and intention, in the sense of having the intention to do something. Intention is only a particular kind of Intentional state, because we always have the intention to do something or to be in a state. But as it will be deeply discussed in the sequel, while belief is an Intentional state, belief is not an intentional mental attitude, in the sense that it is not produced by an intended action or aim at doing something. But it is always an Intentional mental attitude as intention, desire, choice or goal, are.

Intentionality should not be confused either with its close homonym: intensionality (opposed to extensionality). Intensionality is a linguistic concept qualifying some propositions lacking existential generalization and existential substitution properties[1]. Further investigations of the links between intentionality and intensionality were made by analytic philosophers and are out of the scope of this thesis.

We present in the following section the Intentional state that we will refer to all along this dissertation, that is belief, and give an overview of its philosophical ground.

[1] Both following inference principles characterize utterances with the existensionality property:

- **Existential generalization**: From $F(a)$ we infer that $(\exists x), F(x)$. For example for the sentence: "The French king is bald", this principle is not valid because there is no king in France.

- **Existential substitution**: From $F(a)$ and $a = b$ we infer that $F(b)$. For example the inference from "Paul thinks that his father John is kind and John is the hangman" to "Paul thinks the hangman is kind" is not valid because Paul does not know that his father is the hangman.

2.2.2 An important intentional state: belief

This dissertation is mainly focused on the particular Intentional state of belief. First of all, we have to remark that, following philosophical tradition, we will use the word belief in a much stronger sense than in natural language. For example, when we say: "I believe that the sun will shine tomorrow", we express a kind of uncertainty: we are not sure of the future weather. We only express that for me the sun will shine. If we were surer, we would use a stronger expression such as: "I am sure that the sun will shine tomorrow". The philosophical sense of belief is much stronger because it generally implies this uncertainty. It rather refers to our internal view of the world (we will present below a more precise description).

Beyond this idea of belief, no consensus has emerged from different philosophical streams about what it really means to believe and its definition and properties. Among various theories, we present here only an overview of two ones. The first links belief to the agent's mind whereas the second one links it rather to his behavior.

Intuitively, when we try to represent a belief (for example the belief that it is raining outside), we imagine a kind of entity (the belief about a representation[2] of the fact that it is raining outside) lying in our mind, as a data stored in a register of a computer memory. Moreover this belief will influence our behavior: as we believe that it is raining outside, we will carry an umbrella when we leave our home whereas we would take a cap if the sun was sunning. This intuitive view of belief is very close to the *representational* approach to belief: "central cases of belief involve someone's having in her head or mind a representation with the same propositional content as the belief" (Schwitzgebel, 2006)[3].

A diametrically opposite account has been developed by *dispositionalists* with for example (Braithwaite, 1932) and (Marcus, 1990). They consider that an internal representation is not needed to ascribe a belief to an agent. Typically consider the example of an alien, called for instance Clark, adopted by a human family when he was a child. He grows quietly in a small burg and becomes journalist in a big city without anybody knowing that he is an alien. Nothing allows to make some hypotheses about his internal representation capacity and organization, but everybody would ascribe beliefs to him, because his behavior is human like. This idea is the ground of the dispositionalists' view of belief: "Traditional dispositional views of belief assert that for someone to believe some proposition P is for that person to possess one or more particular behavioral dispositions to P" (Schwitzgebel, 2006). The liberal extension of dispositionalism extends this description of belief to avoid the objection that two individuals with same beliefs could act oppositely, by not only reducing analysis of action to a belief but also to desire and other inner feelings (Audi, 1972;

[2]Note that this mental representation needs a language to *code* it: the *language of thought*. The question of the language of thought is far from the scope of this dissertation (interested lectors can read (Fodor, 1985) for more details). In the sequel and in particular in the logical section 3.1, we use propositional modal logic language.

[3]Among other proponents of the representational view of belief, we can cite (Fodor, 1975; Fodor, 1981; Fodor, 1987; Fodor, 1990), (Millikan, 1984; Millikan, 1990) or (Dretske, 1988).

Schwitzgebel, 2002).

Interpretationists concentrate themselves on the concept of observable behavior. They consider that we can ascribe a belief that P to an agent "if its behavior conforms to a pattern that may be effectively captured by taking the intentional stance and attributing the belief that P" (Schwitzgebel, 2006). This theory has been mainly developed by Davidson (Davidson, 1984) and Dennett (Dennett, 1987; Dennett, 1991). For example when we see Clark carrying an umbrella to go outside, I can ascribe to him the belief that it is raining and the desire to keep perfect his brushing, because this ascription helps me to understand his behavior.

Belief is the most important and most studied doxastic state, but some other states pertaining to belief (*i.e.* doxastic states) have been distinguished from belief. We can cite among others: acceptance, holding true (Engel, 1998) or holding as true (Ullmann-Margalit and Avishai, 1992). For example we can hold true without believing: "I may hold true the sentence "all equivalent infinite sets have the same transfinite cardinal" on the authority of a mathematician. But if we do not understand what it says, we cannot believe its content" (Engel, 1998). Similarly we can accept without believing: "I may accept for prudential reasons the proposition that it will rain tomorrow, and act on it, without really believing it". (Meijers, 2002). In the sequel only belief and acceptance will be more deeply studied because these two notions are central in dialogues, as we will show in Section 5.5.

2.2.3 Collective Intentionality

The representational approach is generally the most intuitive and commonly accepted belief account. Thus as a collective or a group does not have a mind, it appears that it is at least metaphorical or fallacious to ascribe a belief to a group. And by extension to other mental attitudes, collective Intentionality is often viewed as *a way of speaking*.

But against this immediate and intuitive idea of individualism, Searle among other authors defends the idea of a genuine collective Intentionality: "the capacity for collective intentionality is biologically innate, and [that] the forms of collective intentionality cannot be eliminated or reduced to to something else" (Searle, 1995, p. 37).

Against this idea of a collective intentionality, Meijers for example answered that a group does not have a so-called mind or consciousness (Meijers, 2002). Collective intentionality appears thus questionable. But Searle does not go as far as to defend the idea of a collective mind: "we can defend the notion of collective intentionality without being "committed to the idea that there exists some Hegelian world spirit, a collective consciousness, or something equally implausible" (Searle, 1995, p. 25)" (Zaibert, 2003).

We can find arguments for collective intentionality in (Tollefsen, 2002). She shows that group also can be viewed as Intentional agents in the same sense as individual agents. She uses the theory of *interpretationist* to defend her point of view: that is "the view that if an agent is interpretable then they

are an intentional agent" (Tollefsen, 2002, p. 88). Interpretationism presented above is not reduced to individual human being but can be applied to any being, and in particular to groups. For example consider the following actual example. Microsoft has attempted a takeover against Yahoo!. We can explain this takeover by ascribing to Microsoft the desire to strengthen its position on the Internet market and the belief that it can buy the company: Microsoft is thus interpretable from the intentional stance. Thus thanks to interpretationism we can ascribe intentionality in general and belief in particular to groups.

In the sequel we will take for granted the existence of a collective intentionality and study more precisely how to ascribe to a group the particular intentional state of belief. We can remark that several works have already defined and characterized some attitudes such as group intention (Tuomela, 2005; Grosz and Sidner, 1990; Cohen and Levesque, 1988) or group emotions (Gilbert, 2002b). But the ascription of a belief to collective seems to be much more polemical, as we will show in the following. Albeit in common language a collective belief is often ascribed to collectives (as in the examples below) the precise nature of this attitude is still under discussions.

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

EXAMPLE. *The United States believes that those responsible for these dreadful acts must be punished. (Gilbert, 2002a, p. 35)*

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

As far as Meijers is concerned (Meijers, 2002), the notion of group belief is a spectrum from an aggregate of individual beliefs ("*opinion poll* conception") to a group belief taken as a whole ("*agreement-based* conception"). In the sequel, we present, analyse and question both bounds of this spectrum.

2.3 Reductionist approaches of group belief

Traditionally, collective belief has rather been viewed as a label of a particular configuration of individual beliefs. They are called "summative" by Quinton (Quinton, 1976) and Gilbert (Gilbert, 1987), "statistical" or "aggregative" by Tuomela (Tuomela, 1992) and described as an "opinion poll conception" by Meijers (Meijers, 2002). The key point of those approaches is that such a collective belief is strongly linked to individual beliefs. They can be reduced to individual beliefs (hence they are called "reductionist approaches") and thus do not exist as a whole. We present in the sequel Gilbert's and Tuomela's reductionist approaches. Gilbert (Gilbert, 1987) exposes two different approaches: a very simple one, based only on individual beliefs and a more complex one using common knowledge, and Tuomela (Tuomela, 1992) introduces one based on the notion of *we-belief*.

First Quinton introduced the term "summative" to describe the lone kind of belief that he considers being ascribable to a group of agents:

> "Groups are said to have beliefs, emotions and attitudes and to take decisions and make promises. But these ways of speaking are plainly metaphorical. To ascribe mental predicates to a group is always an indirect way of ascribing such predicates to its members. With such mental states as beliefs and attitudes the ascriptions are of what we have called a summative kind. To say that the industrial working class is determined to resist anti-trade-union laws is to say that all or most industrial workers are so minded." (Quinton, 1976), from (Hakli, 2006)

In the sequel, we will use the word "summative" to qualify a kind of belief in a broader sense than Quinton's: if a group believes p in a summative sense, this implies that most of its members actually believe it.

2.3.1 Gilbert's simple summative account.

As a first attempt, Gilbert proposes a simple summative account close to the one of Quinton, only based on the notion of personal beliefs and defined as follows:

DEFINITION. *[Simple Summative Account (Gilbert, 1987)] A group I believes that p if and only if most of its members believe that p.*

This account could appear to be well-adapted to capture examples such as:

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

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

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

A first criticism that can be raised against this account is that it does not take into account the group formed by agents. For example, consider a set of agents whose do not know each other, without any link (social, cultural, geographical...) between them. If each agent believes that p, a group belief (in the previous sense) that p holds would be attributed to this set. This property appears to be really too strong.

Moreover in the particular case of the Zuni tribe, its members actually believe that north is the region of force and destruction (in the sequel we will simply write p to represent this sentence), but keep secret this belief, thinking that no other member thinks so, and thus that everyone will mock him if he expressed/revealed it. In this case, it appears that the opinion poll approach is to weak to capture the group belief notion. A group can have such a belief

without any group member being aware of it because agents have no access to other members' private beliefs. Nobody can thus express and defend this belief on behalf of the whole group. For example, no Zuni can discuss with a member of a neighbour tribe and asserts as a kind of representational member of his tribe that Zunis believe that north is the region of force and destruction.

We can extend the above criticism to take into account links between agents of a group: if every agent of the group believes that p and that every other member believes it too, but thinks that they are alone to have this information, this remains to weak to have a kind of group belief. For example if every fierce Zuni warrior thinks that p, and believes that other members think so themselves, without being aware that others are aware that himself believes this sentence, no tribe members will dare to assert that there is a group belief about p. Doing so he thinks that he would reveal his own thoughts.

By iterating these remarks, we come to the conclusion that we need a more complex account with at least the common belief notion to correctly take into account such examples. We can note that a subjective common belief, *i.e.* an agent believes that a common belief on p holds, would neither be strong enough for the same reasons as above because it implies, as beliefs can be wrong, that this common belief could not hold.

2.3.2 Gilbert's complex summative model : the common knowledge account

A more complex account is based on the notion of common knowledge to characterize collective belief. The notion of common knowledge is technically rather complex: there is common knowledge in group I that p if and only if every G's member knows that p and knows that every member knows that p and so on infinitely[4] (another technical characterization is given in terms of the fix-point axiom *cf.* following chapter). This concept helps to capture the social feature needed in group belief definition. It has been technically defined in, among other works, (Lewis, 1969; Lewis, 1972), (Schiffer, 1972) and (Heal, 1978). It has also been introduced in economics in (Aumann, 1976).

Gilbert (Gilbert, 1987) uses Common Knowledge in her complex summative account of the group belief:

DEFINITION.*[Common Knowledge Account (Gilbert, 1987)] A group I believes collectively p iff:*

(1) most of the members of I believe that p, and

(2) it is common knowledge in I that (1).

[4] A closely related notion has been introduced to represent not common knowledge but common belief. It is defined similarly by replacing knowledge by belief in the above definition. Distinctions between these two notions are only related to those between knowledge and belief *i.e.* mainly that believing p does not entail that p is true.

This approach seems more realistic to capture the notion of collective belief. In particular, a group with such a belief would be aware of its own belief. It is a particularly interesting feature of the common knowledge to be public: every member of the group is aware, as in the following example.

EXAMPLE.*[(Tuomela, 1992)] The Finns believe that sauna originated in Finland.*

But this approach is not free from criticisms either. For example, it is too deeply related to individual beliefs. In particular it does not allow members of the group to hold private beliefs distinct from the collective belief, which is however a particularly interesting feature of the proper group belief (as detailed in the following section).

2.3.3 Tuomela's we-belief account

Tuomela (Tuomela, 1992) introduces the "we-belief" account of group belief. He proposes his own aggregative account of group belief defined as a shared we-belief, *i.e.* a we-belief shared by every group member.

More generally, he defines a we-attitude as "a psychological attitude ascribed to a member of a group, say I, that the members denotes by the pronoun "we"" (Tuomela, 1995, p. 37). Indeed a we-attitude is the *internalisation* by an agent of an attitude ascribed to the group of which he is a member. This definition can be applied to any kind of attitude (belief, desire, intention...). It is important to note that a we-attitude does not imply in his theory that this attitude actually holds: an agent can have erroneous beliefs about group mental attitudes.

A we-attitude can be characterized by:

DEFINITION.*[We-attitude (Tuomela, 1995, p. 38)] An agent i has a* we-attitude *to φ in group I iff i*

(1) shares φ in the mode of that attitude,

(2) believes that φ is so shared in I and

(3) believes also that it is mutually believed[5] *in I that φ is so shared in I.*

Thus Tuomela (Tuomela, 1992) uses this *we-attitude* notion to introduce his "simple we-belief approach" of group belief. A group belief expressible by "We believe that p" is a shared we-belief, *i.e.*, single members' we-beliefs that are shared in the group (*cf.* Miller's approaches (Miller, 1990) for a more sophisticated version). This can be characterized by[6]:

DEFINITION.*[Simple We-belief Account] A group I believes φ as "We believe that φ" if and only if every agent i member of I believes*

[5] Note that mutual believe is equivalent to common belief presented later. Tuomela also gives an interesting discussion about the usefulness of infinite conjunctions, in particular he assumes that in many cases only two iterations are necessary (Tuomela, 1995) chapter 1 and (Tuomela, 1984) chapter 7.

[6] Note that we-attitudes are simplified in the case of beliefs.

1) that φ and,

2) that it is mutually believed in I that φ.

For Tuomela, this approach is essential to represent collective beliefs in cases where the set of agents is more an aggregate than a social/structured group, such as:

EXAMPLE.*[(Tuomela, 1992)] The Finns believe that sauna originated in Finland.*

That means that each Finn believes personally that Finland is the country of origin of sauna and that this fact is commonly believed by the Finns. When he considers Finns in the aggregative sense, he denotes only the set of agents with the common feature to have the Finn nationality. No hierarchical link between the Finns is taken into account. But Finns could also be used as a structured group when we consider it with its institutions, its hierarchy between agents[7].

2.3.4 About insufficiencies of reductionist approaches

Although the above approaches seem sufficient to represent some cases of common belief, such as the example of Finns among others, they cannot account for all cases of collective beliefs. Consider the three following examples:

EXAMPLE.*[(Meijers, 2002)] The cabinet believes that genetically modified food is safe and that it should not be forbidden.*

EXAMPLE.*[(Tuomela, 1992)] The Government believes that war against Iraq will begin soon.*

EXAMPLE. *The French government believes it can raise the growth rate to 2,25% in 2007 and 2,5% in 2008.*

The above presented accounts are indeed insufficient to capture these examples. Gilbert (Gilbert, 1987) and Meijers (Meijers, 2002) highlight some arguments against the summative account and sketch thus a more general and complete group belief account.

A first criticism against summative accounts (and the complex summative account in particular) can be built on the following example by (Gilbert, 1987):

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

It seems too strong to ascribe a group belief to this human beings' set that fire burns. The above account does not require a social group formed by agents. Collective belief in this sense is a too casual phenomenon that takes into account

[7] In this case we would use rather the world Finland instead of Finns, as in the sentence: "Finland declares war against United States".

neither the existence of the social group nor its influence on agents. In this case it is not relevant to specify the group of agents because he does not take part in the formation of this belief. Moreover this group does not seem to have any link with the fact that fire burns.

(Meijers, 2002) gives three additional arguments for another kind of group belief account. First, the collective belief has a binding effect on the group members. In Example 2.3.4 every agent *qua* group member of the government has to express and act accordingly to the fact that government believes that war in Iraq is imminent. He should also defend and argue for this belief if it is challenged by another agent, as if it was his personal and private belief[8]. Every agent takes it also for granted that every other government member will act so. He thus cannot change his mind *extempore* and individually, *i.e.* without any discussion with other government members: he is bound to this group belief and changing his mind at the social level should be the result of a group consensus.

Above summative accounts do not take into account this binding feature: indeed nothing in common belief as defined above does have any binding or persistent feature. Moreover as soon as an agent privately changes his mind[9], for any reason or evidence and independently of other group members, the common belief and thus the collective belief should be dropped.

Secondly Meijers argues that this commitment to collective belief is conditioned by its acceptance by other members and their commitment to it. Consider the following example that is closed to the Prisoners Dilemma.

EXAMPLE. *Consider two criminals that were arrested. They claim both that they are innocent. We can thus ascribe them a collective belief that they are innocent. They are examined separately: they still claim that they are innocent and that their partner is also innocent. But if a policeman informs one of them that the other has denounced him then (if he believes the policeman), he would consider that their binding commitment is broken and could denounce his accomplice.*

Thus every group member is committed to defend the group belief in front of anyone, but as soon as one member violates this commitment, other members do not have to defend the collective belief and the group belief cannot hold anymore. This feature, as the above one, cannot be understood in a reductionist account of group belief.

The last and perhaps the more important argument of (Meijers, 2002), also cited by (Gilbert, 1987), against the reductionist approaches is the fact that collective belief should be independent from individual beliefs. In fact, the French government can believe that it can raise the growth rate and every minister can defend this perspective, while some of them are privately convinced that it is infeasible. We can also imagine extreme cases (in this case Tuomela calls the collective belief *spurious* collective belief) where no government member

[8] But this should not induce anything for the private beliefs, only on his behavior.

[9] In Gilbert's complex account, only most of the agents and not all are needed to have a group belief. But the above remains: the idea is that if some agents privately change their mind, the group belief would be modified independently of any discussion and consensus.

thinks privately that the growth can be raised at such a rate. Tuomela gives the following relevant example:

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

This example highlights that the group can believe a statement without any member believing it privately: thus group belief does not imply individual belief. But in most cases, collective and individual beliefs fit together. Example 2.3.1 shows that although every member of the Zuni tribe believes that the north is the region of force and destruction, we cannot ascribe a group belief to the tribe because they keep secret their feeling/belief. We thus can say that individual and collective beliefs are independent.

Durkheim in (Durkheim, 1982) has already expressed that any proper group belief must be "external to individual consciousness". In fact a group belief can be the result of a negotiation, deliberation, persuasion process and thus of a consensus between two or more parts with very different viewpoints. It can even be the result of more or less ethical processes as propaganda or threat as in example 2.3.4. Thus Durkheim's feature allows to handle cases where collective belief is the result of a discussion and where a compromise between each disputant has been reached, as in the following example:

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

With these criticisms as starting point, a new trend appeared, led by Gilbert (Gilbert, 1987), who considers the group belief as a whole, *i.e.* no more reduced to other attitudes.

2.4 Non reductionist approches: toward a proper group belief account

2.4.1 The Plural Subject Account (Gilbert, 1989).

Let us begin by considering the following example from Gilbert:

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

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

In opposition to "summative" approach, Gilbert proposes in her book (Gilbert, 1989) the following characterization of what we can name proper group belief:

DEFINITION.*[The plural subject account (Gilbert, 1989)]*

1. A group I believes that p *if and only if the members of* I *jointly accept that* p.

2. *The members of* I jointly accept *that* p *if and only if it is common knowledge in* I *that the members of* I *individually have intentionally and openly expressed their willingness jointly to accept that* p *with the other members of* I.

As uttered above, the agents' individual beliefs are not taken into account in this formalization. Moreover this definition is by no way related to any individual attitudes. Indeed they always are private, *i.e.* unaccessible to other agents. The only way to access indirectly (and thus with risks of mistakes) to individual attitudes is by agents' behavior and actions, and their interpretation by other agents. For example it is common that agreement is reached in opposition to some members' private attitudes.

We can also note that this kind of group beliefs implies a common belief between every agent of this group belief: indeed as the joint acceptance needs a common knowledge of every agents about their willingness to accept the proposition, we can deduce that every member is aware of the group belief and even that there is common knowledge of it which also implies common belief of the group belief. Whereas in the above account 2.3.2 the summative group belief needed the mutual knowledge on individual beliefs that the proposition p holds, in this account the common knowledge is only about the group belief itself. This common knowledge represents the public feature of the group belief in the following sense: if p is collectively believed then every group member is aware of it and that others are so, too.

It is important to note that Gilbert uses the word *acceptance* in the common use sense. She does not have in mind the philosophical sense, in which acceptance is opposed to belief (see the next section for more details). Tuomela in (Tuomela, 1992) remarks that this definition is circular: the word *joint acceptance* is used in its own definition. In the light of this remark and the above one, Gilbert proposed a slightly different characterization of group belief:

DEFINITION.*[(Gilbert, 2002a)] The members of a population* P *collectively believe that* p *if and only if they are jointly committed to believe that* p *as a body.*

First of all we need to explain the term *jointly committed* (see also (Gilbert, 1996) for more details). A precise discussion about different kinds of commitment will be presented in the sequel (Section 5.4.2). Joint commitment represents a kind of persistence toward a decision taken by group, similarly to personal commitment to stick to an intention until it is fulfilled (Cohen and Levesque, 1990a). As far as Gilbert is concerned, a joint commitment is formed by the expression by every group member of his readiness to be committed with the group. This expression is performed under hypotheses of sound and complete communication, *i.e.* that there is common knowledge in the group of what has been expressed.

2.4.2 A refinement: Tuomela's version of proper collective belief (Tuomela)

Tuomela in (Tuomela, 1992) discusses Gilbert's group belief definition and shows in particular that in the cases of structured groups (for example with representatives) Gilbert's approach (Gilbert, 1987) needs to be adapted. Tuomela takes the following example:

EXAMPLE.*[(Gilbert, 1989)] The United States believe that the [Soviet] invasion of Afghanistan was an unconscionable act.*

Not all Americans accepted this utterance, we only need that a small subgroup called the government, that we could name also the leaders or representatives, accepted it. Tuomela's purpose is not to give an example of a subgroup which imposes its beliefs to the whole group. Every member of the government could believe the opposite of the group belief but decide to accept it as the group belief. On the contrary, the other members of the group could be in accordance with it, and only have to tacitly accept it (because they give their decision willingness to their representatives).

Tuomela gives the following analysis of proper group belief, being inspired by his group intention formalization based on the distinction between "operative" and "non-operative" members (the ones who form the intention and the others who accept it tacitly) and definition of the "right social and normative circumstances" (a kind of institution composed by norms, roles, social rules and tasks...):

DEFINITION.*[(Tuomela, 1992)] The group I believes that p in the social circumstances C if and only if in C there are operative members $A_1, ..., A_m$ of I in respective positions $P_1, ..., P_m$ such that:*

1. *the agents $A_1, ..., A_m$, when they are performing their social tasks in their positions $P_1, ..., P_m$ and due to exercising the relevant authority system of I, (intensionally) jointly accept that p, and because of this exercise of authority system, they ought to continue to accept and positionnally believe*[10] *it;*

[10]Positional beliefs are beliefs that each agent inherits from his function: for example

2. *there is a mutual belief among the operative members* $A_1, ..., A_m$ *to the effect that (1);*

3. *because of (1), the (full-fledged and adequately informed) nonoperative members of I tend tacitly to accept, or a least ought to accept, p, as members of I;*

4. *there is a mutual belief in I to the effect that (3).*

This characterization seems much more realistic and complex than Gilbert's. We can view it as a generalization of the previous approach. It keeps the main aspects introduced by Gilbert: this group belief is not related to individual beliefs, it is public (in the sense presented above) and requires joint acceptance of all members. Moreover, if we consider that every agent is an operative agent and that the set of propositions expressing social and normative circumstances is empty, Tuomela's approach is brought down to Gilbert's one.

Tuomela's view of group belief presented in (Tuomela, 1992) appears as an extension of Gilbert's account. But in contrarily his more recent work (Tuomela, 2000) is more critical in particular with the question of the nature of Gilbert's group belief in the context of the distinction between group belief and group acceptance. This dispute is presented in the following section.

2.4.3 Against Gilbert's plural subject account: the Rejectionnist trend

Some voices raised against Gilbert's plural subject account of group belief, rejecting the fact that this account describes a kind of belief. We can cite among those who are named *Rejectionnists* (Gilbert, 2002a) by Gilbert, K. Brad Wray (Wray, 2001; Wray, 2003), Anthonie Meijers (Meijers, 1999; Meijers, 2002; Meijers, 2003b), Raimo Tuomela (Tuomela, 2000).

First of all it is important to highlight that Rejectionnists mostly do not reject the collective intentionality or the idea to ascribe mental attitudes to a group, neither are they opposed to the plural subject account proposed by Gilbert (Gilbert, 2002a). Thus this position cannot be described as individualism, in the sense that mental attitudes can only be ascribed to individuals, and collective Intentionality can always be reduced to individual Intentionality. The nature of the phenomenon the plural subject account describes is rather the dispute point. Everyone agrees place it into the class of collective doxastic states, but whereas Gilbert argues that it corresponds to a form of belief, K. Brad Wray responds that "... the phenomenon that concerns Gilbert is a species of acceptance [rather than belief]" (Wray, 2001). Thus Rejectionnists agree with non-reductionnist approaches of group belief or more generally of collective Intentionality, but they consider that Gilbert's group belief is not a kind of belief but rather a kind of group acceptance.

"the Flat Earth Society secretary has the positional rule-based belief that the earth is flat" (Tuomela, 1992).

This distinction, although it is interesting in itself for epistemic reasons, allows moreover to describe even more precisely both belief and acceptance notions. Opposite arguments induce the need of a deep and fine analysis of the phenomenon and bring about a better understanding. This helps also to tend toward a precise (and thus formalizable) characterization of group belief.

The distinction between acceptance and belief at the individual layer has already been studied since decades ((Cohen, 1989) among plenty of others), but the key features of each notion are not the object of a consensus between authors yet. This distinction following the key features of each notions presented in Section 2.4.3.1 will be mostly used as basis to the dispute between Gilbert's Believers (Tollefsen, 2003) and the Rejectionnists and as guideline of the current section.

2.4.3.1 Belief versus Acceptance at the individual layer

Whereas belief has been studied for decades (Hintikka, 1962) as representative of doxastic mental states, acceptance has only been examined more recently to study, among others, more precisely the nature of argument premises (Stalnaker, 1984) or to reformulates Moore's paradox (Cohen, 1989). We present here quickly the distinction between both notions[11].

Semantically, a belief that p is a feeling that p is true (Cohen, 1989) [12], whereas acceptance is "a decision to treat p as true in one's utterances and actions" (Hakli, 2006), "the mental state of having a certain policy" (Meijers, 2002) and serve as "background assumptions in deliberation" (Bratman, 1992).

For example, consider two scientists Fox and Dana that are workmates. Fox believes that Aliens exist[13], because he feels convinced by the evidence of their existence. In contrary Dana does not believe in their existence. But she can all the same accept it as a working hypothesis. This acceptance will lead her experiments in a particular way, until either she is in a state where this hypothesis or its consequences are in contradiction with grounded theory (in this case the acceptance will be dropped) or this hypothesis can be proved and thus grounded.

As belief and acceptance seem to be very close, the question of theirs links

[11] Readers can also see (Lehrer, 1990), (Velleman, 2000) and (Frankish, 2004) for important discussion about acceptance.

[12]

> "First then, and very briefly, belief that p is a disposition, when one is attending to issues raised, or items referred to, by the proposition that p, normally to feel it true that p and false that not-p, whether or not one is willing to act, speak, or reason accordingly. But to accept the proposition or rule of inference that p is to treat it as given that p. More precisely, to accept that p is to have or adopt a policy of deeming, positing, or postulating that p – *i.e.* of including that proposition or rule among one's premises for deciding what to do or think in a particular context, whether or not one feels it to be true that p." (Cohen, 1992, p. 4)

[13] Note we really refer to a belief here. It is not a case of faith such as when we say: I believe in God.

appears immediately. Most authors (Tuomela, 2000; Bratman, 1992; Cohen, 1992) consider that both notions are independent, *i.e.* without entailment links. We can accept a proposition without believing it: *cf.* the above example with Dana or similar examples with scientists in (Van Frassen, 1980). We can also believe a proposition without accepting it: we can believe genuinely that our climbing partner has well fixed the rope at his harness, but we will accept it as true only after having checked, by a precautionary principle (similar examples are in (Bratman, 1999)).

In particular, Clarke (Clarke, 1994) argues the "Entailment thesis", *i.e.* that acceptance implies belief (at least to some minimal degree as Tollefsen shows (Tollefsen, 2003)). For example, if Dana were absolutely convinced that aliens do not exist, we cannot understand why and how she could accept even as working hypothesis their existence. At least at the instant of her acceptance she should admit that there is at least a chance that they exist, *i.e.* she believes it at a very small degree. In contrary, Stalnaker (Stalnaker, 1984) has a broader view of acceptance. For him, "to accept a proposition is *to treat it as a true proposition* in one way or another"[14] and thus in his view belief entails acceptance.

Some authors, without any consensus either, have highlighted features that distinguish belief and acceptance. Five main ones emerge (Hakli, 2006; Meijers, 2003b):

- Beliefs are not subject to the agent's will, whereas acceptances are voluntary.
- Beliefs aim at truth, acceptance at utility (they depend on goals).
- Beliefs are shaped by evidence, whereas acceptances need not be.
- Beliefs come in degrees and acceptances are binary.
- Beliefs are context-independent whereas acceptance depends on context.

But in fact, this distinction is not as clear as it is stated. For instance (Hakli, 2006) rather considers that the actual distinction between both doxastic notions is based on voluntarism. (Tuomela, 2000) exhibits some cases of acceptance that aim at truth[15]: for instance consider the Dana example or the one where "a person [...] learns that she is a bunch of leptons and hadrons, and rationally accepts it as true without really starting to believe it" (Hakli, 2006). This

[14]

> "Acceptance, as we shall use this term, is *a broader concept than belief*: it is a generic propositional attitude concept with such notions as presupposing, presuming, postulating, positing, assuming and supposing falling under it. [...] To accept a proposition is *to treat it as a true proposition* in one way or another - to ignore, for the moment at least, the possibility that it is false. [...] To accept a proposition is *to act, in certain respects, as if* one believed it." (Stalnaker, 1984)

[15] Following (Engel, 1998) he even makes a distinction between acceptance as true and pragmatic acceptance.

person has not accepted this physical truth thinking it can bring him some utile counterpart. As a famous professor has taught him, he will accept it trusting in his reputation but he cannot belief it yet. (Hakli, 2006) shows that belief and acceptance can, but do not need to, be shaped by evidences. For example consider a person who is afraid of little dogs. He may believe, without any evidence, that they will bite him if he tries to caress them. That is a kind of instinctive fear. But if he sees that some people have caressed Putty, the chihuahua of her neighbour, he may accept (in a first time) thanks to these evidences that Putty is gentle and try to caress him. Moreover following the traditional modal logic paradigm, we will consider in the sequel that belief and acceptance are only qualitative[16].

In the sequel of this section, we will detail conflicting points opposing *Rejectionnists* and *Believers*, *i.e.* authors defending the plural subject account as being a belief account.

2.4.3.2 The question of the method

In general, the method used by Rejectionnists is to start their argumentation from the distinction presented above between belief and acceptance (and also other doxastic attitudes) at the individual level, to apply it at the collective layer and to find some asymmetries between individual and collective beliefs in their relation with characteristics distinguishing belief and acceptance. They thus conclude that collective belief in the plural subject account sense is not a kind of belief.

But for Gilbert (Gilbert, 2002a), a fallacy appears in this demonstration. She argues that it is unfair to reduce belief only to individual belief. As occurrences of belief can be ascribed to individuals as well as to groups (as the above examples show), features of belief, as general concept, have to integrate attributes of both. For example if the absence of willingness is a key feature of individual belief (allowing to distinguish it from other doxastic attitudes) and that collective belief is produced by the joint willingness of group members, then will should not be considered as a key attribute of belief in general. It is only a contingent tool to discriminate two individual attitudes. Thus it appears fallacious to distinguish group belief and group acceptance on the basis of will and, more generally, of other features that cannot be proved to be constitutive of belief.

But, as K. Brad Wray shows in (Wray, 2003), there is a major weakness in Gilbert's above argument: indeed Gilbert's argument appears difficult to defend. It seems to be intuitive to admit that belief has characteristics of both kinds of belief. But Gilbert's fallacy is to take as granted that the result of the plural subject account is a kind of collective belief. Whereas the questioned point is the nature of her group belief, she asserts that it is a kind of group belief and thus that belief characteristics should be updated in consequence. Using this updated belief, it is unfair to infer that collective belief is a kind of belief.

[16] Note that some works have examined quantitative belief (Laverny and Lang, 2005).

(Meijers, 2003b) gives another kind of criticisms against Gilbert's argument. Gilbert considers important to oppose her top-down methodology (*i.e.* features characterizing belief are extracted from a general notion of belief) to the Rejectionnists' bottom-up methodology (their starting point is rather individual beliefs as basis for belief features). Meijers argues that with a deep enough investigation, we should achieve the same conclusion whatever methodology we used. Moreover he argues that Gilbert's criticism is misguided by the fear of individualism: if individual belief is the starting point of the discussion, she is afraid to have to face individualism *a priori*, *i.e.* that a group belief considered as a whole is a mistake (or is at best metaphorical). In contrary, Rejectionnists would be rather her allies w.r.t. the existence and relevance of collective intentionality. The question of the ascription of a doxastic mental state labeled as "belief" to a group is really the dispute point. Indeed within Rejectionnists themselves there is no consensus about the possibility to ascribe a non-reductionist group belief to some group. On the one hand Meijers (Meijers, 2002) argues that a reductionist group belief can be ascribed to an aggregated group (*i.e.* group without structure) while a non-reductionist cognitive state (result of the plural subject account for example) can be ascribed to a structured group. But these cognitive states are kinds of acceptance; thus no group belief as a whole can be ascribed to a group. On the other hand other Reductionists do not seem to reject the idea of a collective belief at least for some groups; but in their view Gilbert's plural subject account does not describe such a belief (only an acceptance).

It appears thus in the light of this discussion that the Rejectionnists' method is well related and well adapted to discuss the nature of Gilbert's theory method. We detail in the sequel their various arguments opposed to Gilbert's account.

2.4.3.3 Belief and Context

Contrarily to belief, acceptance is often viewed as context-bound. Whereas our beliefs are independent of the context and of the role played in this context, acceptance is deeply related to such social or pragmatic features. For example as a seller of a specific brand of soda, we need to accept that our soda is really the best one and act in accordance in order to sell more soda even if we prefer another brand of soda and thus believe that the latter one is better. Beliefs are not affected by the context in such a way.

(Meijers, 1999) considers that contrarily to individual belief, collective belief in Gilbert's sense is context-dependent and argues that: "we as a group believe that p ... always given a particular situation, role, or point of view. Genuine beliefs, on the other hand, are not context-bound. A person believes that p regardless of context she is in" argues (Meijers, 1999). He argues that individuals are always influenced by the group they belong to, due to the fact that they have a particular role to play in this group. Thus they do not act freely but rather accordingly to their role. Furthermore he keeps on his argument by referring to the example of the cabinet of a genetically modified food company: "The cabinet's belief that genetically modified food is safe and should not be forbidden is typically context bound" (Meijers, 2002). For Meijers it appears

clear that in this case, the belief agreed by group members is typically tied to the particular context of this particular company committee. As group members of a genetically modified food company they have to agree to the safety of these products.

Consider the following example:

EXAMPLE. *[(Gilbert, 1989)] Phil, Cass, Ben and Ted are the members of the campus improvement committee. In a session of the committee they discuss what new amenities might be needed on campus. Each then expresses his readiness to be committed with the other members of the committees to believe as a body that there needs to be a cafe on campus.*

(Gilbert, 2002a) answered that collective belief is not always or necessarily in context. She argues that in the above example of the campus improvement committee, it is "unproblematic" to consider that the committee has a collective belief and that this belief is not tied to a particular context. Indeed as the collective belief is precisely produced by a joint commitment to believe (without any explicit or implicit context), we cannot ascribe a context to the group belief and this collective should have the same relation to context as individual beliefs. She concludes thus that collective belief, as defined, is not always context-bounded. Moreover following her view about the method she argues that if collective belief is nevertheless sometimes bounded to a particular context, this suggests that belief in general is so, too.

Following such arguments, we could also answer to Meijers's example of the company committee: the board of directors does not have this collective belief in a particular situation or context. Whatever the context in which the cabinet will be (or more precisely any of its member will be), he will defend this belief: for example in front of a governmental commission, in an ecological meeting or with farmers, any representative of the cabinet will stay close to the cabinet's collective belief. For the cabinet, this statement publicly represents the real world. In our view the confusion comes from the fact that collective belief emerges from the expressed acceptance of each cabinet member in the particular context of the cabinet (*i.e.* acceptance *qua* group member). Each member accepts that genetically modified food is safe for a particular reason, for personal pragmatical considerations or because he actually believes that this statement is true. After the joint agreement, the cabinet as a whole believes that genetically modified food is safe whatever the context is in which the cabinet will be, without pragmatic considerations.

(Meijers, 2003b) agrees with Gilbert's example showing that group belief is not indeed context-related in some cases. He also agrees with the fact that individual or collective beliefs can be context-bound. But, although he concedes that collective belief is not always context-bound, he gives two examples in which he thinks that collective belief is deeply linked to a particular context whereas individual belief is not. Firstly, "in order for a belief to be a collective belief there has to be a *shared* understanding of the context of such a belief among the members of the group" (Meijers, 2003b). For example, in the above example of the cabinet, the exact sense of the terms "genetically modified food"

needs to be discussed and grounded among cabinet members. Thus the collective belief is tied to a particular context in which the sense of the context has been grounded. Meijers remarks that this need of grounding does not exist for individual belief. We can firstly answer to Meijers that this shared understanding is socially grounded privately and locally in the group; it is independent of the context, the institution in which the group is. Although this consensus on a shared understanding is not needed in an individual, we have to remark that an individual understanding of the believed proposition is nevertheless needed. We can only accept that we are a bunch of leptons and hadrons (*cf.* above Hakli's example) but we need at least to understand what it means to really start to believe it. Individual and collective beliefs appear thus similar on this point.

"Secondly, collective beliefs are context bound in the sense that they are *role* bound" (Meijers, 2003b). This can be illustrated by the cabinet example: whatever the cabinet members individually believe, in their role of group members they believe that genetically modified food is safe and can be consumed. Individuals' beliefs are not so linked to roles. But Meijers is once again "guilty of committing the fallacy of composition" (Wray, 2003): he still considers collective belief at the sub-level of its members and not at the layer of the group as a whole. In the cabinet example, positional beliefs in Tuomela's sense (Tuomela, 1992) or simply members acceptance rather than genuine beliefs are taken into consideration.

Rejecting Meijers' arguments we stay close to Gilbert's view on this point: collective belief is no more context-related than individual beliefs; they can both have a link with context in some particular cases whereas acceptance is context-bound. We now discuss the links between collective belief and evidence.

2.4.3.4 Belief and Evidence

Belief is often viewed as a feeling that works it's way into our mind led by evidence supporting it, whereas acceptance is rather driven by pragmatic or prudential reasons. For example, as a lawyer we can trust in the innocence of our client but accept his guilt in order to plead guilty in front of judges (because we think that we risk too much by pleading his innocence). Thus link to evidence is viewed as a key feature to distinguish belief from acceptance.

Rejectionnists argue on the one hand that "responsiveness to evidence distinguishes belief from acceptance" (Wray, 2003) and in the other hand that belief formed by a group is formed only on pragmatic/prudential criteria. Thus this group attitude should not be named belief but acceptance.

For (Meijers, 1999), "we may give up our ... collective beliefs for reasons not related to the epistemic evidence we have but, for example, because of prudential concerns". The following example illustrates this point:

EXAMPLE.*[(Meijers, 1999)] An military alliance of states jointly believes a bombing campaign against a foreign country will achieve a certain result. But the members of the alliance of states fear the political repercussions at home of*

this particular bombing campaign, "and consequently the alliance may give up its belief" that the bombing will achieve the desirable result.

Indeed (Meijers, 2002) argues that similarly to acceptance, collective beliefs are not solely shaped by evidence; practical reasons have also an important role. In this presentation of the example, the goal of protecting their own country becomes stronger than their collective belief, thus this pragmatic reason seem to have led them to drop their former belief.

Gilbert (Gilbert, 2002a) gives two objections against this example. Firstly, although the individual members beliefs of having an important benefit in case of a giving up of the group belief and the fear to maintain this collective belief obviously play *some role* in the abandon of the group belief, but nothing grounds clearly that they are the key reasons to give up the collective belief.

Moreover, and this is deeply linked to previous objection, we can imagine a common scenario in which this group belief is dropped for purely epistemic reasons (Gilbert, 2002a): consider this alliance composed of three states represented by Peter, Antoine and Karl, having the collective belief that bombing will achieve their goal. But each having his own reason, all want to give up this belief. We can imagine that Karl speaks first and questions the truth of this belief: "Is the object of our belief really going to achieve our goal?". As other members want to give up the collective belief, they will seize the opportunity and utter that it is not going to achieve the goal. The alliance will reach a joint agreement to drop the pre-existent belief (because every member have agreed to the truth of the sentence: "Bombing will not achieve our goal") and to adopt the belief that bombing will not achieve their goal on the basis of epistemic reasons[17]. Thus Gilbert have shown that purely epistemic reasons can also be the reasons to drop this belief.

But of course Karl could also have questioned the interest for the alliance to keep or give up this belief and reached the same conclusion as previously but for prudential reasons. But as Gilbert showed, human beings can also assess their belief in the light of both epistemic and prudential considerations. Consider the example of Joe and his unfaithful wife:

EXAMPLE. *Joe may wonder if it is in his interests to believe that his wife is unfaithful. (Gilbert, 2002a)*

Pragmatic considerations will drive here Joe's behavior: Joe may decide not to belief his wife is unfaithful, else the belief would make him miserable.

Gilbert argues in (Gilbert, 2002a) that Rejectionnists are wrong because "responsiveness to evidence" does not allow to distinguish collective belief from individual one and that both individual and collective beliefs can be dropped for epistemic and pragmatic reasons.

(Wray, 2003)'s point is that the distinguishing factor is not the link between doxastic states and evidence but rather between doxastic states and goals of the agent: agents accept a view in light of their goal whereas they cannot

[17] Of course Antoine and Peter could have answered Karl by expressing their preference to give up the collective belief, but it would have been irrelevant here.

believe something for the same reason[18]. Moreover, contrary to what Gilbert expressed, Meijers (Meijers, 2003b) and he recognize that beliefs can sometimes be produced for non-evidential reasons. As Hume explains in (Hume, 1977) "it is a miracle that so many people persist in believing something that is so contrary to experience". People can be sometimes so impervious to evidences that we must accept that belief cannot be only directed by evidence and that something else influences our beliefs. But it would be too strong to conclude that people can believe what they choose, as the man of Gilbert (Gilbert, 2002a) that can choose between believing that his wife is unfaithful and that she is not. Wray argues that this influence (of this "something else") is in fact "as much out of our control as the impact that evidential considerations have on us" and that Gilbert's example would be better formalized with acceptance: the man can accept or refuse to accept at will the fact that his wife is unfaithful (but it would be a kind of Hume's miracle that the husband does not believe his wife unfaithful with some evidence under his eyes).

(Meijers, 2003b) questions both Gilbert's answers to Rejectionnists' criticisms about links between group belief and evidence. In the one hand she argues that collective belief may be given up for epistemic reasons (*cf.* the example of the alliance) and in the other hand that personal belief may be dropped given practical considerations (*cf.* the example of the deceived husband). Firstly Meijers clarifies his previous argument: he did not assume that collective belief cannot be dropped for epistemic reasons, but rather that it is not *solely* shaped by evidence, because the procedure needed to create, alter or drop a collective belief is "never purely epistemic" (Meijers, 2003b). Secondly, Gilbert's example of the deceived husband is highly questionable for Meijers too, and in particular he sees in this example a case of acceptance rather than one of belief. It seems more intuitive to represent this example as: Joe chooses to accept that his wife is unfaithful and to act according to this acceptance, otherwise he would be miserable.

But (Hakli, 2006) questions all arguments about the link between belief, acceptance and goal for and against plural subject account as a account of proper group belief, because he thinks that the question is hedged by some common preliminary hypotheses: acceptance is necessarily goal-dependent and this feature is a key distinguishing factor. He argues that, as (Tuomela, 2000) highlights, there are kinds of acceptance that are only dependent on epistemic goals and thus that their distinction from belief on this criteria appears problematic. Moreover he argues following (Molden and Higgins, 2005) that goals, interests or emotions can interact in the belief formation process: for example, it appears that we have a inclination to be more sensitive to evidence in favor of our beliefs. Thus for him goal-dependency does not distinguish between belief and acceptance, and he argues that the discussion should be shifted toward another feature.

We can illustrate his purpose by following examples. Indeed, some individ-

[18] Moreover as groups are typically constituted by their goals, every group belief is in part shaped by goals.

	Gilbert	Meijers	Wray	Hakli
Belief	E/P	E	E/P	E/P
Group belief	E/P	E/P	P	E/P
Acceptance	P	E/P	P	E/P

Table 2.1: Summary of positions

ual beliefs can be adopted in the light of prudential considerations (see Joe's example) whereas some views can be accepted following evidence of their truth. For example if Joe surprises his wife with her lover, he will have to accept her two-timing due to evidence (or if he stays unrealistic he can deny evidence and still trust in his wife's faithfulness). Moreover as Meijers and Wray consider that "the purpose of accepting is to advance some goal", it appears obvious that if a view has evidence for its truth, these evidence will have big influence on its acceptance. Furthermore with Joe's and alliance examples, she has showed that both individual belief can be dropped for practical reasons and collective belief for epistemic ones. We can thus conclude that the link with evidence does not allow to discriminate belief and acceptance.

We can summarize this section by adding to Hakli's table (Table 1 (Hakli, 2006) his own position on the table 2.1. In this table, "E" means that the attitude is affected necessarily by epistemic factors, "P" means that the attitude is affected necessarily by pragmatic factors and "E/P" means that the attitude is necessarily (or can only be) affected by both kinds of reasons.

We will follow Hakli's position and shift this discussion toward another feature, that is indeed closely related to both former studied ones.

2.4.3.5 Belief and Truth

Belief is often viewed as aiming at truth whereas acceptance aims at utility, because it depends on goals. For example we can illustrate this point on the above example of the lawyer: a lawyer trust in the innocence of his client because of evidence means that for him it is true that his client is innocent but he will plead guilty only for utilitarian reasons. We can see immediately that this discussion is deeply linked to both previous features.

(Gilbert, 2002a) assumes only that due to previous considerations about evidence and context, collective belief hold the same relation with truth as individual belief.

For (Meijers, 2002), at least three points indicate that collective belief contrarily to individual belief aims does not aim at truth: collective belief is context-bound, collective belief can be drop for prudential reasons and we may collective belief that no member takes to be true. He defends this point with the selection committee example. As every member aims at truth and that the collective belief resulting from their consensus is different from their individual beliefs, we can conclude that the aim of this belief is not at truth but rather follows the goal of the committee: the candidate does not need to really be the best but

only the one who is the object of the consensus.

Moreover he adds in (Meijers, 2003b) that this issue is deeply related to the two following ones: the role of epistemic considerations in the group belief and the link between the group belief and group members' individual beliefs. If only epistemic considerations shape collective beliefs, *i.e.* if the aim of the truth is the only point of the collective belief, Meijers do not understand why a commitment towards other members is needed to build a collective belief. If collective beliefs relied to the truth of its content, why does this belief need to bind agents which others? Moreover, Meijers argues that if the creation of the collective belief is led by evidence, there should not be independence between collective and individual belief. He supports thus the argument that a group cannot create a proper group belief following the plural subject account. A group can thus only have a collective belief in the "opinion poll conception" and a collective agreement-based acceptance when the set of agents is viewed as a structured group of agents. As previously we can still reproach to Meijers to fall into the "fallacy of composition" by linking deeply collective and individual beliefs.

As this part of the discussion is closely related to previous ones, he does not bring us some factors to answer our question. I will discuss in the following section the link between belief and will that promises more debates.

2.4.3.6 Belief and Will

Traditionally, in most comparisons between belief and acceptance (see (Meijers, 1999), (Engel, 1998), (Wray, 2001) and (Williams, 1970) for example), a key distinction between these two doxastic states is their relation to will: it is claimed that we cannot believe something at will whereas acceptance can be viewed as the result of a kind of intentional action.

EXAMPLE. *I cannot believe at will that I shall never die (Meijers, 2002).*

EXAMPLE. *I accept some proposition for the sake of an argument by a voluntary and intensional mental action or an expressed act of concession.*

For Meijers, collective beliefs "... require some sort of voluntary assent, agreement, or decision by the members of the group for the belief." (Meijers, 1999, p. 64). Members decide voluntarily to agree to be jointly committed with other members. Thus for (Meijers, 1999) and (Wray, 2001), as Gilbert's collective belief is necessarily at will, the obvious conclusion is that collective belief is not a kind of belief but rather a kind of acceptance.

(Gilbert, 2002a) questions the premise of the above argument. She argues that there is a confusion about the agent that wills. Indeed the collective belief needs not to be willed by the group, as a whole. Of course collective belief is produced voluntarily by the agreement of each individual agents to be jointly committed to believe as a body, but this body has not the will to have this belief. Wray concedes this point to Gilbert and adds that "Meijers is guilty of committing the fallacy of composition" (Wray, 2003), *i.e.* when he refers to the

belief of a group, he indeed refers to a composition of individual members' belief and not to the one of the group as a whole.

But can a collective belief be produced only by the will of a collective? Suppose that a group adopted a collective goal to have the belief that p (by agreement of each member to be jointly committed with each other to have this belief). Indeed the group did not create the collective belief automatically and each member still needs to accept to be jointly committed with each others to this belief. We can thus argue that a collective, as an individual agent, cannot believe at will; the hypothesis of involuntarism[19] still holds.

Meijers handles this issue only by answering that there is a "different ontological orientation" (Meijers, 2003b) between him and Gilbert. Indeed the fact to distinguish the will of the group as a whole and the will of each members is problematic for him (Meijers, 2003a).

(Hakli, 2006) has a slightly different view upon this question. For him, voluntarity is the feature that distinguishes acceptance among other doxastic states. But his definition of voluntarity is a bit different from the common one.

DEFINITION. *[Voluntarity] An attitude that p is* voluntary *(V) for an agent if and only if the agent has direct control over the attitude in the sense that the agent can directly adopt, revise and abandon the attitude either at will or by forming an intention. The complement property of voluntarity is involuntarity ($\overline{V}$), which means that it is not possible directly to adopt, revise and abandon the attitude at will. (Hakli, 2006)*

For Hakli, acceptances are those doxastic states that can be in principle created, changed or dropped at will. Thus he argues that collective belief is a kind of acceptance because, if the group wants to change one of its beliefs, he can do it, whatever the sense given to the group is.

(Gilbert, 2002a) has opposed the following argument:

EXAMPLE. *Fran and her friend Trudy explicitly adopt as their collective goal their collectively believing that their future is bright, without any concern for whether or not their future actually is bright.(Gilbert, 2002a)*

She argues that the key point of this example is that, after having adopted their goal, their group belief is not created automatically and immediately only thanks to the former goal agreement. Group members need to discuss and to jointly accept this belief. Thus for Gilbert, "a group cannot directly bring its belief into being by an act of its own will". Hakli answers that collective belief seems nevertheless to fit with his notion of voluntarity since after having decided to have this particular goal the group does not need anything else as additional evidence to adopt its collective belief[20]. We can note that the temporal aspect of

[19] For Gilbert, this word refers to the fact that "We cannot bring a belief of mine into being by act of will, or not directly" (Gilbert, 2002a)

[20] Hakli remarks that in some specific cases that he excludes for the sake of demonstration, group belief cannot be grounded due to conflicts or inconsistency with *e.g.* constitutive rule of the group.

Gilbert's example does not appear here. Hakli can thus conclude that collective belief is a kind of acceptance.

But to answer to Hakli's argument, we need to make precise the sense of the words "direct control" in his definition of voluntarity. As far as we are concerned, and by referring to the above argumentation, the direct control of an agent over an attitude refers to his capability to create, change or drop it at will, *i.e.* by an intensional action, without the inference of other agents and with his own capacity. Thus we cannot strictly say that the group has a direct control over his beliefs because he needs a discussion and a consensus among his members to create his belief. A group needs, and we think that is the sense of Gilbert's example, that his members jointly commit to believe in order that the group believes.

Moreover (Gilbert, 2002a) and (Tollefsen, 2003) question the fact that voluntarism is a discriminatory feature between belief and acceptance. An agent has nevertheless a kind of power over his beliefs. He has the ability to place himself in a particular situation to obtain a specific belief. For example (Tollefsen, 2003), in order to write a dissertation student reads some articles and books and not other ones and thus by this choice between articles he places himself favorably to gain some beliefs. Thus belief is for Gilbert between voluntarism and involuntarism. Moreover if we accept with (Tollefsen, 2003) that belief has at least a little part of voluntarism, it appears that belief is also goal-directed. For example we can refer again to Tollefsen's argument. The student has gained a lot of beliefs in order to write his dissertation and with the goal to become PhD.

Once again no argument emerges against the plural subject account of group belief. But the question of the link between belief and will is disputable, in particular because we consider the will of the group and its capacity to create a belief. The question of the will of a group and his link with members' will is out of the scope of this dissertation. We stay close to Gilbert's view on this point. And we consider group will in a non-reductionist paradigm. In this context we feel that Gilbert's arguments is quite acceptable and no salient distinction appears between individual and collective belief w.r.t. will.

Besides this main argumentation, we can cite some additional arguments in favor of the being of a genuine group belief described by Gilbert's plural subject account.

2.4.3.7 Additional arguments

The centrality of belief. Tollefsen (Tollefsen, 2003) reuses Davidson's "centrality of belief" principle (Davidson, 1984). In its weakest version, it says that an agent cannot have some propositional attitude without belief. A stronger version states that all intentional states get their content from the web of beliefs. That does not mean an agent needs to believe a proposition p in order to have the intention that it becomes true. The centrality of belief argues only that the agent needs to have some beliefs around this proposition to intend it.

Tollefsen's argument is that, due to the principle of the Centrality of belief

and to the Rejectionnists point that a group can have acceptance but cannot have belief, Rejectionnists have to abandon either collective intentionality or Rejectionnism (Tollefsen, 2003). A group cannot thus have an acceptance in the non-reductionist sense without having a belief in the same sense. We note that this argument cannot be uses to defend Gilbert's group belief account but only to defend the existence of a genuine group belief.

The entailment thesis. Tollefsen (Tollefsen, 2003) uses against Rejectionnists Clarke's "Entailment Thesis" (Clarke, 1994). Clarke argues that an agent cannot accept a proposition without believing it and thus that acceptance is just a kind of belief. Tollefsen applies this argument to the collective layer to reject the argument that a collective cannot have a kind of belief but only acceptance.

She defends this thesis by arguing that to accept some proposition as true, an agent has to believe it at least at a certain degree. We do not give more details about this argument, because it uses a notion of belief that gets a bit to far from the one introduced in the beginning of this chapter.

Collective belief and collective acceptance. Along this section, we have presented the defense of Gilbert's plural subject account of collective belief against Rejectionnists who see it rather as a kind of acceptance. This defense can be more powerful by representing collective acceptance in parallel with collective belief.

For instance let us consider Gilbert's example:

EXAMPLE. *[(Gilbert, 2002a)] The campus improvement committee is discussing whether the campus needs a cafe. Cass says, "Look, let's suppose, for the sake of argument, that Café Sunshine – the only cafe in town – is going to close in the next year." Phil says, "Okay", and the others nod in agreement.*

It seems that similarly to the collective belief created by the the plural subject account, members of the group have created a collective acceptance. This collective attitude seems indeed to have really different features from collective belief. Its study and its link with individual acceptance (is this collective acceptance a kind of acceptance, or is it something else ?) is out of the scope of this thesis. This collective acceptance distinct from collective belief is an additional argument in favor of the plural subject account of Gilbert.

In this section we have deeply discussed the nature of the result of Gilbert's plural subject account. Contrarily to the Rejectionnists point of view, we consider that this discussion has shown that it is a kind of group belief. Preparing the future logical formalization of group belief we highlight in the next section key features that characterize group belief.

2.5 Toward a formal characterization of group belief

In the following chapter, we will propose a formalization of group belief as a modal logic operator. Before this, we need to highlight the main aspects that can be extracted from the characterization in the present chapter, which will guide the subsequent formalization. Moreover as it has been shown in the previous section, proper group belief is very close to individual belief. Formal properties have thus also to be close. The three following ones can be viewed as proper features of collective beliefs. They also distinguish genuine group belief from aggregate group belief (that will be represented in the sequel as a common belief in the sense defined above). Of course this characterization is based on Gilbert's plural subject account.

2.5.1 Proper group belief is in no case related to individual beliefs

This property is likely the major criticism against the summative approaches. Moreover, inspired by Durkheim (Durkheim, 1982), it is the most important contribution of Gilbert's account.

It means that there will be no entailment link between our group belief operator and the individual belief operator. Thus our group belief operator will be able to take into account "spurious beliefs" (Tuomela, 1992) *i.e.* cases of group beliefs, as that in the Communist Party of Ruritania example. Contrarily common belief is deeply linked to individual beliefs: common belief implies and is defined from individual beliefs.

2.5.2 There is a kind of commitment on proper group belief

As soon as the group belief has been established, even if some group members disagree with this belief, they must act in compliance with it, *i.e.* they are committed in some way to this belief. When they violate it, they are liable for sanctions, ranking from blames of the group (Gilbert, 1987) to exclusion from the group (Tuomela, 1992). In the sequel we will be stronger by logically forbidding violation or inconsistency.

2.5.3 The group members share a mutual belief about proper group beliefs

One of the major criticisms against the "simple summative approach" of group belief is that every group member can believe individually that φ without any collective belief on φ because agents are not aware of what other agents believe. A kind of mutual belief is thus necessary, but not about the content of the group belief (as in the "complex summative approach"), but rather on the group belief

itself. Tuomela (Tuomela, 1992) defends this thesis arguing that group belief is grounded due to a joint and intentional group action.

This feature characterizes the fact that the group belief is public. This means that every group member is aware of the truth or falsity of all groups beliefs. As we will show in the following chapter, common belief does not share this property: whereas there is common belief that a common belief holds, there is not common belief that a common belief does not hold, due to the fact that beliefs can be wrong.

2.6 Conclusion

This chapter has been the place to introduce and explain theoretical concepts necessary for the rest of this dissertation. In particular we have introduced intentional states for individual agents and defended their ascription to a group of agents. Afterwards we have shifted the question to the more precise question of the nature of belief. Existing accounts of group belief have been presented and discussed. In particular we have followed Gilbert's plural subject account of group belief because it well represents group belief: not as an aggregation of individual beliefs but as a whole. This account has been defended against the Rejectionnists trend. They consider indeed that this collective belief is rather a kind of acceptance. It has been shown that group belief is closer to individual belief than acceptance. To prepare the next chapter, we have highlighted key features that we will need to formalize in our group belief logical formalization. In particular it should not be related to individual beliefs and should be public for group members.

In the sequel (Chapter 3) our logical framework will include individual and collective belief operators but also a common belief operator. At this point we stay close with authors like Tuomela, who argues that "group beliefs require a combination of the positional and the aggregative approach" (Tuomela, 1992), or Meijers. For the latter collective belief is rather a continuum between opinion poll conception: "we, the individuals in the group, *as individuals*, believe that p" and the agreement-based conception: "we, the group as a whole, believe that p" (Meijers, 2003b). But contrarily to them, we consider that group belief can be defined only with a proper group belief operator. Indeed as the individual belief will be considered in the sequel as a conscious attitude, we consider that group belief needs also to be conscious, or to be more precise the group needs to be aware of his beliefs. And only the agreement based account has this feature. We will extend this logical framework in Chapter 4 to group acceptance in a institutional context. Group can thus have various acceptance depending on the context.

Chapter 3

The logic of group belief

In this chapter we present a logical formalization of group belief that has the properties presented in the previous chapter. We integrate the group belief logic in a BDI (Belief Desire Intention) framework to link it with other mental attitudes such as choice and intention or with actions. Our logic is based on the logic of belief, choice and action developed in (Herzig and Longin, 2004) which builds on the works of Cohen & Levesque (Cohen and Levesque, 1990a) and Sadek (Sadek, 1992). We augment this logic by a modal operator expressing the group belief. We provide a Kripke semantics (Kripke, 1963) and the associated axiomatics.

In the case of a group reduced to a singleton $\{i\}$, we will identify the group belief operator with the classical individual belief operator *à la* Hintikka (Hintikka, 1962). Thus a particular individual belief operator is superfluous. We introduce also a mutual belief operator to exhibit some logical features linking group belief (in the non-reductionist sense) and an aggregative group belief.

3.1 Syntax

Let $AGT = \{i, j, \ldots\}$ be a finite set of agents. A *group of agents* (or a *group* for short) is a nonempty subset of AGT. We use $I, J, K, \ldots$ to denote groups. When $I' \subseteq I$ we say that I' is a subgroup of I. Let $ATM = \{p, q, \ldots\}$ be the set of atomic formulas. Complex formulas are denoted by $\varphi, \psi \ldots$ Let $ACT = \{\alpha, \beta, \ldots\}$ be the set of actions. We suppose that some actions in ACT are of the form $i{:}\alpha$, where i is the author of (*i.e.*, performs) the action α.[1] We also consider that actions are performed in front of a group of observers, *i.e.* agents that are aware of the performance of the action (*e.g.* hearers in the case of speech act). We omit them in the notation for the sake of clarity.

[1] In particular for a speech act α, this notation allows to specify the author of α without mentioning the addressee, the illocutionary force...

The language of our logic is defined by the following BNF grammar[2]:

$$\varphi ::= p \mid \neg\varphi \mid \varphi \vee \varphi \mid G_I\,\varphi \mid \mathit{Choice}_i\,\varphi \mid \mathit{MBel}_I\,\varphi \\ \mid \mathit{After}_\alpha\,\varphi \mid \mathit{Before}_\alpha\,\varphi \mid \Box\varphi \mid H\varphi$$

where p ranges over ATM, α over ACT, i over AGT, and I over $2^{AGT} \setminus \{\emptyset\}$. Operators are defined in detail in the next section; we give only here a reading of these operators: $G_I\,\varphi$ reads "the group I believes that φ holds"; $\mathit{Choice}_i\,\varphi$ reads "agent i chooses (prefers) that φ"; $\mathit{MBel}_I\,\varphi$ reads "agents in I mutually believe that φ"; $\mathit{After}_\alpha\,\varphi$ reads "φ will be true after α"; $\mathit{Before}_\alpha\,\varphi$ reads "φ was true before α"; $\Box\varphi$ reads "φ is true from now on"; $H\varphi$ reads "φ was true up to now".

The classical boolean connectives $\wedge$, $\rightarrow$, $\leftrightarrow$, $\top$ (tautology) and $\bot$ (contradiction) are defined from $\vee$ and $\neg$ in the usual manner. The operators *Bel* (individual belief) and *Intend* (intention) will be defined as abbreviations.

3.2 Semantics

A model includes a set of possible worlds W and a mapping $\mathscr{V} : W \longrightarrow (ATM \longrightarrow \{0,1\})$ associating a valuation $\mathscr{V}_w$ to every $w \in W$. Models moreover contain mappings that will be detailed in the sequel. As usual, mappings are identified to accessibility relations: for $\mathscr{R}$ a mapping and $w, w' \in W$, $w\mathscr{R}w'$ iff $w' \in \mathscr{R}(w)$.

3.2.1 Group Belief

To each possible world w and each non-empty $I \subseteq AGT$, we associate the set of possible worlds that are consistent with all propositions believed in world w by the group I. This set is characterized by the mapping: $\mathscr{G} : 2^{AGT} \longrightarrow (W \longrightarrow 2^W)$ associating an accessibility relation to each non-empty subgroup of AGT. $\mathscr{G}_I(w)$ contains those worlds where all propositions that are collectively believed hold.

$G_I\,\varphi$ reads "the group I believes that φ is true" or "φ is collectively believed by the group I". When I is a singleton, $G_{\{i\}}\,\varphi$ is identified with the standard belief operator Bel_i *à la* Hintikka (Hintikka, 1962) and thus reads "agent i believes individually that φ holds". We write $G_i\,\varphi$ for $G_{\{i\}}\,\varphi$.

The truth condition for G_I stipulates that φ is collectively believed at w, noted $w \Vdash G_I\,\varphi$, if and only if φ holds in every world that is consistent with the set of collectively believed propositions:

$$w \Vdash G_I\,\varphi \text{ iff } w' \Vdash \varphi \text{ for every } w' \in \mathscr{G}_I(w).$$

We assume that:

❶ $\mathscr{G}_I$ is serial.

[2] Operators are defined in detail in the next section.

Thus collective belief is rational: if a proposition holds in every world that is consistent with the set of collectively believed propositions, then at least one such a world exists.

Furthermore we postulate the following constraints on accessibility relations, for groups I and I' such that $I' \subseteq I$:

❷ if $u\mathscr{G}_{I'}v$ and $v\mathscr{G}_I w$ then $u\mathscr{G}_I w$;

❸ if $u\mathscr{G}_{I'}v$ and $u\mathscr{G}_I w$ then $v\mathscr{G}_I w$;

❹ if $u\mathscr{G}_I v$ and $v\mathscr{G}_{I'}w_1$ then there is w_2 such that $u\mathscr{G}_I w_2$ and

- $V(w_1) = V(w_2)$,
- $\mathscr{G}_K(w_1) = \mathscr{G}_K(w_2)$ for all K such that $K \cap I = \emptyset$,
- $\mathscr{C}_k(w_1) = \mathscr{C}_k(w_2)$ for all k such that $k \notin I$, where $\mathscr{C}$ is the accessibility relation for choice to be defined below;

❺ $\mathscr{G}_I \subseteq \bigcup_{i \in I} \mathscr{G}_I \circ \mathscr{G}_i$.

Constraint ❷ stipulates that agents of a subset I' of the set I are aware of what is collectively believed in the group I: whenever w is a world for which it is believed by I' that all I-believed propositions hold in w, then all I-believed propositions indeed hold in w. This is a kind of *attention* property: each subgroup is aware of what is believed by the group.

Similarly ❸ expresses that subgroups are aware of what is not believed in the group, too.

❷ and ❸ together make that if $u\mathscr{G}_{I'}v$ then $\mathscr{G}_I(u) = \mathscr{G}_I(v)$, *i.e.* if $u\mathscr{G}_{I'}v$ then what is believed by I at u is the same as what is believed by I at v. From ❷ and ❸ it also follows that $\mathscr{G}_I$ is transitive and euclidian.

❹ says that if an information "about something outside group I" (see the definition in the following subsection) is believed by I then it is believed by I this information is believed by every subgroup of I.

❺ says that if it is believed by a set I that a proposition is believed by every agent then it is believed by I, too.

3.2.2 Mutual belief

From individual belief, we define the notion of mutual belief of a group of agents, it thus corresponds to an aggregative notion of group belief. Semantically we have the mapping $\mathscr{MB} : 2^{AGT} \longrightarrow (W \longrightarrow 2^W)$ associating an accessibility relation $\mathscr{MB}_I$ to each group $I \subseteq AGT$. $\mathscr{MB}_I(w)$ denotes the set of possible worlds compatible with mutual beliefs of the group I. For each group I, $\mathscr{MB}_I$ is defined as the transitive closure of the set of accessibility relations associated to the I's members beliefs (*i.e.* $\mathscr{G}_i$ for each $i \in I$):

❻ $\mathscr{MB}_I = (\bigcup_{i \in I} \mathscr{G}_i)^+$

$MBel_I\,\varphi$ reads "it is mutual belief for the group I that φ is true". It means that every member of the group believes individually that φ and that it is mutual belief for the group that φ is true[3]. (See (Fagin et al., 1995) for more details about the logic of mutual belief.)

3.2.3 Choice

Among all the worlds in $\mathscr{G}_i(w)$ that are possible for agent i, there are some that i prefers. Semantically, these worlds are identified by yet another mapping $\mathscr{C} : AGT \longrightarrow (W \longrightarrow 2^W)$ associating an accessibility relation $\mathscr{C}_i$ to each $i \in AGT$. $\mathscr{C}_i(w)$ denote the set of worlds the agent i prefers.

$Choice_i\,\varphi$ reads "agent i chooses that φ". Choice can be viewed as a preference operator and we sometimes also say that "i *prefers* that φ". Note that we only consider individual choices, group choices being beyond the scope of this dissertation.

The truth condition for $Choice_i$ stipulates that $w \Vdash Choice_i\,\varphi$ if φ holds in all chosen worlds:

$$w \Vdash Choice_i\,\varphi \text{ iff } w' \Vdash \varphi \text{ for every } w' \in \mathscr{C}_i(w).$$

We assume that:

❼ $\mathscr{C}_i$ is serial, transitive, and euclidian.[4]

We refer to (Herzig and Longin, 2004) for more details about the logic of choice, and the definition of intention from choice.

3.2.4 Choice and belief

As said above, an agent only chooses worlds he considers possible (see Figure 3.1):

❽ $\mathscr{C}_i(w) \subseteq \mathscr{G}_i(w)$.

Hence what is believed by an agent must be chosen by him (this represents trivial cases of choices in opposition to choices that are not believed), and choice is a mental attitude that is logically weaker than belief.

We moreover require that worlds chosen by i are also chosen from i's "believed worlds", and *vice versa*.

❾ if $w\mathscr{G}_i w'$ then $\mathscr{C}_i(w) = \mathscr{C}_i(w')$.

This constraint means that agent i is aware of his choices.

[3] This definition is recursive. We can define mutual belief with an infinite conjunction: every agent believes φ, that the other agents believe φ, that the other agents believe that every agent believes φ...

[4] Contrarily to the semantics of belief, there is no consensus to the choice operator. In particular it refers to what Cohen and Levesque named "goal" with stronger properties (Cohen and Levesque only assumed seriality). A more detailed comparison with other choice operators is developed in (Herzig and Longin, 2004).

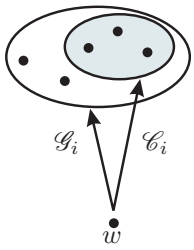

Figure 3.1: Grounding and Choice

3.2.5 Action and time

The model contains a mapping $\mathscr{A} : ACT \longrightarrow (W \longrightarrow 2^W)$ associating an accessibility relation $\mathscr{A}_\alpha$ to every $\alpha \in ACT$. $\mathscr{A}_\alpha(w)$ is the set of worlds accessible from w through the execution of α.

The formula $After_\alpha\,\varphi$ reads: "φ holds after the execution of α". As there is at most one possible execution of α, which is imposed by following temporal constraints, the dual operator $Happens_\alpha\,\varphi \stackrel{def}{=} \neg After_\alpha\,\neg\varphi$ reads: "α is happening and φ is true just afterwards". Hence $After_\alpha\,\bot$ expresses that α does not happen, and $Happens_\alpha\,\top$ that α happens. We often write $Happens\,(\alpha)$ for $Happens_\alpha\,\top$.

The truth condition is:

$$w \Vdash After_\alpha\,\varphi \text{ iff } w' \Vdash \varphi \text{ for every } w' \in \mathscr{A}_\alpha(w)$$

The formula $Before_\alpha\,\varphi$ reads: "φ holds before every execution of α". The dual $Done_\alpha\,\varphi \stackrel{def}{=} \neg Before_\alpha\,\neg\varphi$ expresses that the action α has been performed before which φ held. Hence $Done_\alpha\,\top$ reads: "α has just happened".

The accessibility relation for $Before_\alpha$ is the converse of the above relation $\mathscr{A}_\alpha$. The truth condition is thus:

$$w \Vdash Before_\alpha\,\varphi \text{ iff } w' \Vdash \varphi \text{ for every } w' \in \mathscr{A}_\alpha^{-1}(w).$$

To speak about temporal sequences, the model contains a mapping $\mathscr{R}_\Box : W \longrightarrow 2^W$. $\mathscr{R}_\Box(w)$ is the set of worlds representing future history from w.

The associated modal operator $\Box\varphi$ expresses that henceforth φ holds. A dual operator $\Diamond$ is defined by $\Diamond\varphi \stackrel{def}{=} \neg\Box\neg\varphi$ (meaning that eventually φ holds).

The truth condition is:

$$w \Vdash \Box\varphi \text{ iff } w' \Vdash \varphi \text{ for every } w' \in \mathscr{R}_\Box(w)$$

$H\varphi$ expresses that φ has always held in the past. A dual operator P is defined by $P\varphi \stackrel{def}{=} \neg H\neg\varphi$ (meaning that at one instant in the past φ held).

The accessibility relation for H is the converse of the above relation $\mathscr{R}_\Box$. The truth condition is:

$$w \Vdash H\varphi \text{ iff } w' \Vdash \varphi \text{ for every } w' \in \mathscr{R}_\Box^{-1}(w).$$

Just as Cohen and Levesque, we impose additional constraints on time, action and links between both notions:

① if $w\mathscr{R}_\alpha w'$ and $w\mathscr{R}_\beta w''$ then $w' = w''$;

② $\mathscr{R}_\Box$ is reflexive[5], transitive[6] and linear [7];

③ if $w\mathscr{R}_\alpha w'$ then $w\mathscr{R}_\Box w'$;

④ if $w\mathscr{R}_\alpha w'$, $w\mathscr{R}_\Box w''$ and $w \neq w''$ then $w'\mathscr{R}_\Box w''$.

Constraint ① expresses that, whenever the action performed, the resulting world will be the same which imposes determinism of actions (take $\alpha = \beta$ for example).

Constraint ② expresses that the present is included in the future, that the future of the future of a world is still a future world and finally that the time is linear which entails that there exists an order relation over worlds in relation with time.

Constraint ③ expresses an intuitive relation between action and time: worlds resulting from an action are in the future of the world which means that the performance of an action takes time.

Finally constraint ④ expresses a similar concept: a world in the future of the current state is also a world describing the future of any world after the performance of an action in the current world.

As said above, we do not detail here the relationship between action and individual mental attitudes (belief and choice) and refer the reader to (Herzig and Longin, 2004). We only consider here the link between action and group belief.

3.2.6 Action and group belief

Differently from the Public Announcement Logic (PAL) (van Ditmarsch, van der Hoek, and Kooi, 2005) where announcement (or speech acts for us) are perceived by every agents, we will later only consider dialogical actions and in this case (*cf.* following part for more details about this point), we consider in this dissertation that actions are public for attending agents, in the sense that their occurrences are completely and soundly perceived by every agent. This hypothesis is necessary to ensure the public feature of group belief without using heavy processes such as grounding (see (Traum, 1999) for a description of this process).

For example, when agent i performs an assertive speech act only towards agent j then j will perceive the assertion. If no other agent perceives this action then the attendees are just $K = \{i, j\}$, and the action is public for exactly this group. But when agent i performs a speech act towards agent j in front of an assistance L then the set of attendees is extended to $K = \{i, j\} \cup L$. Agents

[5] For every $w \in W, w\mathscr{R}_\Box w$.
[6] If $w_1\mathscr{R}_\Box w_2\mathscr{R}_\Box w_3$ then $w_1\mathscr{R}_\Box w_3$.
[7] If $w_1\mathscr{R}_\Box w_2$ and $w_1\mathscr{R}_\Box w_2$ then $w_2\mathscr{R}_\Box w_3$ or $w_3\mathscr{R}_\Box w_2$.

outside group K do not change their beliefs. Actions are thus public inside the group of attentive hearers, but remains private in this group considering other agents.

Let α be an action performed by agent i in front of attendees K (of which i is a member). The property of public actions (for group K) corresponds to the constraint:

⑤ $\mathscr{A}_\alpha^{-1}(w) = \emptyset$ if and only if $(\mathscr{G}_K \circ \mathscr{A}_\alpha^{-1})(w) = \emptyset$

This constraint means that an action α has been performed in front of K if and only if K is mutually aware of this performance.

3.2.7 Validity and logical consequence

DEFINITION. *φ is true in M iff $M, w \models \varphi$ for every $w \subset W$.*

DEFINITION. *φ is valid in a class of models C (noted $\models_{\mathsf{C}} \varphi$) iff $M \models \varphi$ for every $M \in \mathsf{C}$.*

DEFINITION. *$S \models_{\mathsf{C}} \varphi$ iff for every $M \in \mathsf{C}$, if $M \models \psi$ for every $\psi \in S$ then $M \models \varphi$.*

3.3 Axiomatics

3.3.1 Group Belief

The logic of the group belief operator is a normal modal logic of type KD:

$$\frac{\varphi}{G_I\,\varphi} \qquad (\mathrm{RN}_{G_I})$$

$$G_I\,(\varphi \to \psi) \to (G_I\,\varphi \to G_I\,\psi) \qquad (\mathrm{K}_{G_I})$$

$$G_I\,\varphi \to \neg G_I\,\neg\varphi \qquad (\mathrm{D}_{G_I})$$

(RN_{G_I}) and (K_{G_I}) are the bases of every normal modal logic and in particular for logic using a possible-worlds semantics. (RN_{G_I}) expresses that every tautology is collectively believed and (K_{G_I}) means that if it is collectively believed by a group I that φ implies ψ then the group belief of φ implies the one of *psi*. (D_{G_I}) expresses that information collectively believed by a group are consistent: it cannot be the case that both φ and $\neg\varphi$ are simultaneously grounded.

In accordance with the preceding semantic conditions the following logical axioms respectively correspond to the constraints ❷ and ❸. Thus, for each $I' \subseteq I$:

$$G_I\,\varphi \to G_{I'}\,G_I\,\varphi \qquad (\mathrm{SR}+)$$

$$\neg G_I\,\varphi \to G_{I'}\,\neg G_I\,\varphi \qquad (\mathrm{SR}-)$$

The axioms of strong rationality (SR+) and (SR−) express that if a proposition φ is believed (resp. not believed) by group I then it is believed by each subgroup that φ is believed (resp. not believed) by I. This is due to the public character of the group belief operator.[8]

Corresponding to the semantic constraint ❺, we have the following axiom of group belief:

$$(\bigwedge_{i \in I} G_I\, G_i\, \varphi) \rightarrow G_I\, \varphi \qquad \text{(CG)}$$

It expresses that if it is collectively believed by I that every member of I individually believes φ, then it is collectively believed by I.

It is important to remark that we will give a particular status to the formula $G_I\, G_i\, \varphi$ (with $i \in I$). Literally it means that the group I believes that i believes individually that φ holds. Indeed we will consider that this formula has as primary origin the expression by the agent i of his belief that φ holds. Following speech act theory with (Searle, 1969), the assertion of φ by agent i counts as the public expression of his belief that φ holds. It is afterwards accepted as a group belief by the group. We can notice that other group members do not have any access to the truth of the formula $Bel_i\, \varphi$. Thus unless the acceptance of this fact induces an inconsistence with previous group beliefs, we can consider as a shortcut that this expressed individual believed is automatically grounded in the group (a detailed account will be given in Chapter 5 about speech acts). Note that this does not imply that individual agents actually believe that i has this belief, but this represents the fact that he has expressed it publicly. Thus we consider for the moment without additional details that $G_I\, Bel_i\, \varphi$ represents (the effect of) the expression by agent i of his own belief.

The last axiom must be restricted to particular formulas, viz. objective formulas for a group, that we define as follows.

Definition. The set of formulas that are objective for a group I is defined inductively to be the smallest set such that:

- every atomic formula p is objective for I;
- $G_K\, \varphi$ is objective for I if $K \cap I = \emptyset$, for every formula φ;
- $Choice_j\, \varphi$ is objective for I if $j \notin I$, for every formula φ;
- if φ and φ' are objective for I then $\neg\varphi$, $\varphi \wedge \varphi'$ are objective for I.

[8] In particular (SR+) and (SR−) axioms are a generalization to a group of the (positive and negative) introspection axioms commonly accepted for mental attitudes (like belief, choice...): each agent i member of the group I if aware of what is believed (resp. not believed) by the group I:

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

With respect to the semantic constraint ❹, our fourth axiom, a weak rationality axiom, stipulates that if I' is a subgroup of I and φ is objective for I then:

$$G_I\,\varphi \rightarrow G_I\,G_{I'}\,\varphi \tag{WR}$$

(WR) expresses that if φ is objective for group I and believed by I then it is necessarily believed by I that the formula is believed by each subgroup I'.

Note that this does not imply that φ is actually believed by every subgroup, *i.e.* (WR) does not entail $G_I\,\varphi \rightarrow G_{I'}\,\varphi$. In particular, the fact that φ is believed by group I does not imply that the members of I believe that φ.

It is very important to note that (WR) concerns only formulas φ that are objective for I. Indeed, if we applied (WR) to some mental states of an agent of the group, we would restrict the agents' autonomy.

Now if agent i asserts that $G_j\,\varphi$ in presence of group I, then the formula $G_I\,G_i\,G_j\,\varphi$ holds afterwards, and if (WR) applied unrestrictedly then j could not express later that he ignores whether φ, or believes $\neg\varphi$. If he made this last speech act, the formulas $G_I\,G_j\,\neg\varphi$ and, thanks to (WR), $G_I\,G_i\,G_j\,\neg\varphi$ would hold, which is inconsistent with the above formula $G_I\,G_i\,G_j\,\varphi$ (Gaudou, Herzig, and Longin, 2006b).

Together, (WR) and (CG) stipulate that for formulas φ that are objective for I we have:

$$(\bigwedge_{i\in I} G_I\,G_i\,\varphi) \leftrightarrow G_I\,\varphi \tag{3.1}$$

Thus in the case of objective formulas φ, it is collectively believed that φ holds if and only if it is believed by the group I that every member believes it. This theorem represents a kind of establishment process of the group belief by consensus.

From axioms (SR+) and (SR−), we can prove that we have the modal axioms (4) and (5) for G_I operators as theorems of our logic:

$$G_I\,\varphi \rightarrow G_I\,G_I\,\varphi \tag{4_{G_I}}$$
$$\neg G_I\,\varphi \rightarrow G_I\,\neg G_I\,\varphi \tag{5_{G_I}}$$

Thus operator G_I is in a normal modal logic of type KD45. Hence for individual belief we obtain its standard KD45 logic.

We can moreover show that if $I' \subseteq I$ then:

$$G_I\,\varphi \leftrightarrow G_{I'}\,G_I\,\varphi \tag{3.2}$$
$$\neg G_I\,\varphi \leftrightarrow G_{I'}\,\neg G_I\,\varphi \tag{3.3}$$

These theorems express that subgroups of a group are aware of what is believed (resp. not believed) in the group. The formula $(\bigwedge_{I'\subseteq I} G_I\,G_{I'}\,\varphi) \rightarrow G_I\,\varphi$ is provable from our axiom (CG).

Moreover we can prove that:

$$G_I\,\varphi \leftrightarrow G_I\,G_{I'}\,G_I\,\varphi \tag{3.4}$$
$$\neg G_I\,\varphi \leftrightarrow G_I\,G_{I'}\,\neg G_I\,\varphi \tag{3.5}$$

These theorems say that if φ is (not) believed by a group, then it is believed by this group that it is believed by every subgroup of this group that φ is (not) believed for the group.

Even if I' is a subgroup of I we do not necessarily have $G_I\,\varphi \rightarrow G_{I'}\,\varphi$. Such a principle would be too strong because it would restrict the autonomy of subgroups I' of I: a proposition can be believed by a group I while there is a dissident subgroup I' of I, *i.e.* a group which believes the contrary: $G_I\,\varphi \wedge \neg G_{I'}\,\varphi$ is consistent in our logic even if $I \cap I' \neq \emptyset$.

3.3.2 Mutual belief

Axiomatically mutual belief is defined by the Fixpoint Axiom, which expresses that a mutual belief about φ holds if and only if every agent believes that φ and that the mutual belief holds (see (Fagin et al., 1995) for more details):

$$MBel_I\,\varphi \leftrightarrow \bigwedge_{i\in I} G_i\,(\varphi \wedge MBel_I\,\varphi) \qquad (\mathrm{FP}_{MBel_I})$$

and the Least Fixpoint axiom, that will be used in the sequel:

$$\bigwedge_{i\in I} G_i\,\varphi \wedge MBel_I\,(\varphi \rightarrow \bigwedge_{i\in I} G_i\,\varphi) \rightarrow MBel_I\,\varphi \qquad (\mathrm{LFP}_{MBel_I})$$

From this axiomatics, we can deduce that $MBel_I$ is a normal modal operator of type KD4:

$$MBel_I\,\varphi \rightarrow \neg MBel_I\,\neg\varphi \qquad (\mathrm{D}_{MBel_I})$$
$$MBel_I\,\varphi \rightarrow MBel_I\,MBel_I\,\varphi \qquad (4_{MBel_I})$$

Note that the negative introspection axiom 5 is not a theorem of the $MBel$ logic: $\neg MBel_I\,\varphi \rightarrow MBel_I\,\neg MBel_I\,\varphi$ does not hold. In particular from (D_{MBel_I}) and (5_{MBel_I}) we could deduce that $MBel_I\,MBel_I\,\varphi \rightarrow MBel_I\,\varphi$ holds. This formula is to strong in the case of the mutual belief, indeed as belief is by definition fallible, every agent can be wrong in their belief: for example $\bigwedge_{i\in I} G_i\,MBel_I\,\varphi$ can hold whereas $MBel_I$ does not.

3.3.3 Mutual belief and group belief

By definition, it comes from the Fixpoint Axiom that when it is mutual belief for a group I that φ then necessarily every member of I believes individually that φ:

$$MBel_I\,\varphi \rightarrow \bigwedge_{i\in I} G_i\,\varphi \tag{3.6}$$

Now we will show the link between mutual belief of a group I and group belief of the whole group I:

THEOREM. *We have the equivalence:*

$$G_I\,\varphi \leftrightarrow MBel_I\,G_I\,\varphi \tag{3.7}$$

That means that a formula φ is believed by a group I if and only if there is mutual belief in the group that φ is collectively believed. This property is mainly due to the public nature of the group belief operator.

PROOF.

1. $\vdash MBel_I\,G_I\,\varphi \rightarrow G_i\,G_I\,\varphi$, *by theorem (3.6);*
2. $\vdash MBel_I\,G_I\,\varphi \rightarrow G_I\,\varphi$, *from 1. by theorem (3.2);*
3. $\vdash G_I\,\varphi \rightarrow G_i\,G_I\,\varphi$, *by theorem (3.2), for every* $i \in I$*;*
4. $\vdash G_I\,\varphi \rightarrow \bigwedge_{i\in I} G_i\,G_I\,\varphi$, *from 3. because it holds for every* $i \in I$*;*
5. $\vdash MBel_I\,(G_I\,\varphi \rightarrow \bigwedge_{i\in I} G_i\,G_I\,\varphi)$, *from 4. by the Rule of Necessitation for* $MBel_I$ *;*
6. $\vdash MBel_I\,(G_I\,\varphi \rightarrow \bigwedge_{i\in I} G_i\,G_I\,\varphi) \rightarrow (\bigwedge_{i\in I} G_i\,G_I\,\varphi \rightarrow MBel_I\,G_I\,\varphi)$, *from axiom* LFP_{MBel_I} *;*
7. $\vdash \bigwedge_{i\in I} G_i\,G_I\,\varphi \rightarrow MBel_I\,G_I\,\varphi$, *from 5. and 6. by Modus Ponens;*
8. $\vdash G_I\,\varphi \rightarrow MBel_I\,G_I\,\varphi$, *from 4. and 7.;*
9. $\vdash G_I\,\varphi \leftrightarrow MBel_I\,G_I\,\varphi$, *from 2. and 8.*

□

We can also highlight that contrarily to mutual belief, axiom 5 holds for group belief. This is due to the fact that group belief is public (in the sense that agents are soundly and completely aware of grounded group beliefs) whereas the notion of mutual belief is a completely public notion but not soundly so, because beliefs can be wrong.

3.3.4 Choice and intention

With respect to the semantic constraint ❼, the choice operator is defined in a normal modal logic of type KD45 and we have the axioms (D_{Choice_i}), (4_{Choice_i}) and (5_{Choice_i}):

$$Choice_i\,\varphi \rightarrow \neg Choice_i\,\neg\varphi \qquad (\mathrm{D}_{Choice_i})$$
$$Choice_i\,\varphi \rightarrow Choice_i\,Choice_i\,\varphi \qquad (4_{Choice_i})$$
$$\neg Choice_i\,\varphi \rightarrow Choice_i\,\neg Choice_i\,\varphi \qquad (5_{Choice_i})$$

(D_{Choice_i}) means that an agent cannot have inconsistent choices whereas (4_{Choice_i}) (resp. (5_{Choice_i})) expresses that he choices (resp. does not choice) are preferred states of affairs.

We define intention in a way similar to Cohen and Levesque as:

$$Intend_i\,\varphi \stackrel{def}{=} Choice_i \Diamond G_i\,\varphi \wedge \neg G_i\,\varphi \wedge \neg G_i \Diamond G_i\,\varphi \qquad (\mathrm{Def}_{Intend_i})$$

where $\Diamond$ is an operator of linear temporal logic LTL. Hence i intends that φ if and only if in i's preferred worlds i will believe φ at some world in the future, i does not believe φ holds now (*i.e.* φ is an achievement goal), and it is not the case that i believes he will come to believe φ anyway (φ is not self-realizing). For more details on this intention operator, see (Herzig and Longin, 2004). We often write $Intend_i\,\alpha$ for $Intend_i\,Done_\alpha\,\top$.

3.3.5 Choice and belief

Due to the semantic constraint ❽ we have the following Bridge Axiom:

$$G_i\,\varphi \rightarrow Choice_i\,\varphi \qquad (\mathrm{BA1}_{G_i,Choice_i})$$

which means that every formula grounded for agent i must necessarily be chosen by this agent.

Our semantics also validates the principles:

$$Choice_i\,\varphi \leftrightarrow G_i\,Choice_i\,\varphi \qquad (\mathrm{BA2}_{G_i,Choice_i})$$
$$\neg Choice_i\,\varphi \leftrightarrow G_i\,\neg Choice_i\,\varphi \qquad (\mathrm{BA3}_{G_i,Choice_i})$$

that correspond with constraint ❾. This expresses that agents are aware of their choices.

Moreover by definition of the intention operator, introspection properties also hold for intention, and intention on a formula always implies that belief of this formula does not hold:

$$Intend_i\,\varphi \leftrightarrow G_i\,Intend_i\,\varphi \qquad (\mathrm{BA1}_{G_i,Intend_i})$$
$$\neg Intend_i\,\varphi \leftrightarrow G_i\,\neg Intend_i\,\varphi \qquad (\mathrm{BA2}_{G_i,Intend_i})$$
$$Intend_i\,\varphi \rightarrow \neg G_i\,\varphi \qquad (\mathrm{BA3}_{G_i,Intend_i})$$

3.3.6 Action and time

With respect to the semantic constraints, the action operators $After_\alpha$ and its converse $Before_\alpha$ are defined in a K_t logic, *i.e.* a normal modal logic with following conversion axioms:

$$\varphi \rightarrow After_\alpha\,Done_\alpha\,\varphi \qquad (\mathrm{I}_{After_\alpha,Done_\alpha})$$
$$\varphi \rightarrow Before_\alpha\,Happens_\alpha\,\varphi \qquad (\mathrm{I}_{Before_\alpha,Happens_\alpha})$$

These axioms characterize the fact that the relation $\mathscr{A}_\alpha^{-1}$ is the converse of $\mathscr{A}_\alpha$.

With respect to the semantic constraints, the temporal operators $\Box$ and its converse H are defined in a S4.3 logic, *i.e.* a normal modal logic with the following conversion axioms:

$$\Box\varphi \rightarrow \varphi \qquad (\mathrm{T}_\Box)$$
$$\Box\varphi \rightarrow \Box\Box\varphi \qquad (4_\Box)$$
$$\Diamond\varphi_1 \wedge \Diamond\varphi_2 \rightarrow (\Diamond(\varphi_1 \wedge \Diamond\varphi_2) \vee \Diamond(\Diamond\varphi_1 \wedge \varphi_2)) \qquad (\mathrm{Linear}_\Box)$$
$$\varphi \rightarrow \Box P\varphi \qquad (\mathrm{I}_{\Box,P})$$
$$\varphi \rightarrow H\Diamond\varphi \qquad (\mathrm{I}_{H,\Diamond})$$

Moreover semantics constraints ①, ③ and ④ impose following axioms:

$$\Box\varphi \rightarrow \mathit{After}_\alpha\,\varphi \qquad (\mathrm{Inc}_{\mathit{After}_\alpha})$$
$$\mathit{Happens}_\alpha\,\varphi \rightarrow \mathit{After}_\beta\,\varphi \qquad (\mathrm{Hist}_1)$$
$$\Diamond\varphi \rightarrow (\varphi \vee \mathit{After}_\alpha\,\Diamond\varphi) \qquad (\mathrm{Hist}_2)$$

($\mathrm{Inc}_{\mathit{After}_\alpha}$) means that, if in every instant in the future φ will hold, then φ will hold after the performance of any action. (Hist_1) expresses determinism: if there exists an execution of an action α after which φ holds, then after any performance of any action β, φ will hold. Axiom (Hist_2) expresses that if φ will be eventually true in the future, then either φ is currently true or it will be eventually true after every performance of any action α.

3.3.7 Action and group belief

As we have said above we only consider public actions and α be an action performed by a agent i in front of attentive group I (of which i is member). Thus we have following axiom of public actions corresponding to the semantic constraint ⑤, for each group I observing an action α:

$$G_I\,\mathit{Done}_\alpha\,\top \leftrightarrow \mathit{Done}_\alpha\,\top \qquad (\mathrm{PA}_{I,\alpha})$$
$$G_I\,\neg\mathit{Done}_\alpha\,\top \leftrightarrow \neg\mathit{Done}_\alpha\,\top \qquad (\mathrm{NA}_{I,\alpha})$$

To sum it up, an action has been (resp. has not been) performed by an member of group I if and only if it is believed by the group that it has been (resp. has not been) performed.

3.4 Completeness and soundness of the logic

All axioms are Sahlqvist (Sahlqvist, 1975) except axiom WR; so we have completeness for the class of models satisfying constraints ❶ – ❸, ❺ – ❾ and ① – ⑤.

We did not formally prove completeness for the whole class of models satisfying constraints ❶ – ❾ and ① – ⑤. We conjecture that this can be done by using the filtration method.

3.5 Action laws

Action laws describe the semantics of each action. They 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 be executable. The effects (or postconditions) are properties that hold after the action. 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 7.2. The general form of an executability law is

$$Choice_i\ Happens\ (i{:}\alpha) \wedge Precond(i{:}\alpha) \leftrightarrow Happens\ (i{:}\alpha) \qquad (\mathrm{Int}_{Choice_i,\alpha_i})$$

This expresses a principle of intentional action: an action happens exactly when its preconditions hold and its author chooses it to happen (Lorini, Herzig, and Castelfranchi, 2006; Lorini and Herzig, 2008). The general form of an effect law is $\varphi \rightarrow After_\alpha\ Postcond(\alpha)$. 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 (Fitting, 1983). We have for example:

$$LAWS \models G_i\ After_\beta\ After_\alpha\ Postcond(\alpha)$$

3.6 Example

To highlight our proposal for the semantics of group belief we will describe our running example that we have presented in the introduction. We recall it here. There are three agents $AGT = \{0, 1, 2\}$:

1. Agent 0 (privately) believes that 2 is smart, formally written $G_0\ smart_2$.
2. In private conversation agent 0 tells 1 that 2 is not smart. The illocutionary effect is $G_{\{0,1\}}\ G_0 \neg\, smart_2$.
3. After 1 publicly adopts $\neg\, smart_2$ (*e.g.* by confirming publicly that $\neg smart_2$) we obtain $G_{\{0,1\}}\ \neg smart_2$, by consensus between 0 and 1.
4. When agent 2 joins the conversation, and 0 informs 1 and 2 that 2 is smart: the illocutionary effect is $G_{\{0,1,2\}}\ G_0\ smart_2$.

5. When both 1 and 2 publicly adopt $smart_2$ we moreover obtain $G_{\{0,1,2\}}\, smart_2$.

This illustrates that even for nested groups $J_0 = \{0\} \subset J_1 = \{0,1\} \subset J_2 = \{0,1,2\}$ we might have states of public groundedness for the different groups which are about propositions that are mutually inconsistent, viz. here:

$$G_{J_0}\ smart_2$$
$$G_{J_1}\ \neg smart_2$$
$$G_{J_2}\ smart_2$$

3.7 Back to the philosophical origin

In this last section of the logical formalization of the group belief notion, we can go back to our starting point to show that our formalism stayed close to features highlighted in the previous chapter.

3.7.1 Group belief features

3.7.1.1 Proper group belief is in no case related to individual beliefs

As said above, our group belief operator G_I is linked in no way to the private beliefs, and in particular for every agent i, member or not of the group I, neither $G_I\,\varphi \rightarrow G_i\,\varphi$ nor $G_i\,\varphi \rightarrow G_I\,\varphi$ is a theorem of our logic: neither group belief implies individual belief, nor private individual belief implies anything at the public layer. Nevertheless $G_I\,\varphi \wedge G_i\,\varphi$ is a consistent formula, *i.e.* proper group belief does not ban the existence of a mutual belief. Links between individual and collective beliefs will be introduced by studying additional constraints on agents such as sincerity or cooperation.

3.7.1.2 There is a kind of commitment on the proper group belief

This property is a theorem of our logic: it corresponds to axioms (D_{G_I}) and (WR). If it is collectively believed by a group that φ holds, then no group member can express that he believes $\neg\varphi$ (in the case of φ objective, *i.e.* when the grounding arises from a discussion between every agent and not only from assertion of one member of the group). Indeed, if the group I believes that φ holds, then due to (WR), in particular $G_I\,G_i\,\varphi$ holds for every member i of the group I and with the axiom (D_{G_I}), $\neg G_I\,\neg G_i\,\varphi$ holds, too. No member of I can then perform an assertive speech act with propositional content $\neg\varphi$ (Gaudou, Herzig, and Longin, 2006a).

3.7.1.3 The group members share a mutual belief about proper group beliefs

This feature is a theorem of our logic: as proven above, the formula $G_I\,\varphi \rightarrow MBel_I\,G_I\,\varphi$ is a theorem. Our logic is stronger because we even have the equiv-

alence.

After this first examination our group belief operator has features required in the previous chapter for a proper group belief operator. We will examine more deeply its link with Gilbert's and Tuomela's approaches.

3.7.2 Philosophical account and formal representation

3.7.2.1 Gilbert's plural subject account

In the sequel, we show that our group belief operator is close to the group belief definition given by Gilbert (Gilbert, 1989) thanks to axioms (WR) and (CG). In particular in the case of formula φ objective for I, we have the equivalence:

$$G_I\,\varphi \leftrightarrow (\bigwedge_{i\in I} G_I\,G_i\,\varphi) \tag{3.8}$$

Moreover from the above formula and (3.7), we can deduce the equivalence:

$$G_I\,\varphi \leftrightarrow MBel_I\,(\bigwedge_{i\in I} G_I\,G_i\,\varphi), \text{ for } \varphi \text{ objective for } I \tag{3.9}$$

This equivalence is very close to Gilbert's characterization of group belief. In fact, formula $G_I\,G_i\,\varphi$ typically results from that agent i expressing in front of group I that he believes φ (and this fact has been collectively accepted as a group belief).

Thanks to the axiom (CG): $(\bigwedge_{i\in I} G_I\,G_i\,\varphi) \rightarrow G_I\,\varphi$, by making public his belief that φ holds, agent i expresses also implicitly his acceptance that φ to be grounded in the group. In particular, because he is aware of this theorem (thanks to Rule of Necessitation of the group belief operator), he knows that by expressing publicly that he believes φ, $G_I\,\varphi$ could hold if the other agents do the same.

Thus formula (3.9) can be read : "an objective formula φ is collectively believed by group I if and only if it is mutual belief in group I that every group member publicly expressed that they believe the objective formula φ".

It follows from this informal proof that for objective formulas our group belief operator matches Gilbert's definition. Thus our operator is an admissible formalization of Gilbert's account of group belief (for objective formulas).

3.7.2.2 Tuomela's account

With our simple framework, we obviously cannot formalize the whole complexity of Tuomela's definition. In particular, we introduce neither notions of roles, institution or norms, nor do we distinguish between operative and non-operative agents. But following Tuomela, we will simplify a bit his formalism to show an interesting property. We will ignore the social and normative circumstances (assuming that they do not impose any restrictions).

As said above, the formula $G_I\,G_i\,\varphi$ means: it is believed by the group I that i believes φ. But another reading that we have previously given is: agent i has

expressed in front of I that he believes φ. This formula is in fact the effect of an assert speech act in the sense that $LAWS \models After_\alpha\ G_I\ G_i\ \varphi$ where α is i's speech act of asserting that φ, and $LAWS$ is an appropriate set of action laws for α (such laws will be given in chapters 6 and 7). Thus some non-objective formulas can be immediately grounded as soon as an agent performs the corresponding speech act. These formulas neither require discussion not explicit acceptance by every agent of the group. Only a single agent is needed to ground such a formula for the whole group. It appears clearly that formulas such as $G_I\ G_i\ \varphi$ do not correspond to group belief *à la* Gilbert. We will show that they correspond in fact to a group belief in Tuomela's approach.

In our logic, an agent is operative (in Tuomela's sense) in what concerns his own beliefs: $G_I\ G_i\ \varphi$ holds if and only if agent i, that we can call operative agent in this case, asserts φ. This implies in our logic that he ought to continue to accept publicly that he believes φ. By Theorem (3.7), there is mutual belief that $G_I\ G_i\ \varphi$. The other agents (called non-operative in this case) ought to accept that it is grounded that agent i believes φ. Moreover there is mutual belief about this formula. Due to what we have said above about Tuomela's group belief, we have that $G_I\ G_i\ \varphi$ implies group belief that i believes φ (in Tuomela's sense).

We could be closer to Tuomela's approach by introducing the concept of *leaders* of a group (about a proposition) in our framework, noted $leaders(I, \varphi)$. The leaders would be a subgroup of the group of agents I, verifying properties such as: $G_I\ G_{leaders(I,\varphi)}\ \varphi \leftrightarrow G_I\ \varphi$. This means that if it is grounded for the whole group I that it is grounded for leaders that φ, then φ is *de facto* grounded for the whole group (*i.e.* if leaders have jointly accepted φ, then other agents have to accept it tacitly and thus φ becomes a proper group belief *à la* Tuomela). The group of leaders could be for example the government for every decision concerning the whole nation (thus leaders get their power from citizen's votes) or a group of specialists of a domain for every fact concerning their competence domain (they are leaders due to their knowledge and skills).

3.8 Conclusion

In this section and following the philosophical study presented in the preceding chapter, we have introduced a logical formalization of group belief in a BDI logic and given its semantics and axiomatics, introducing bridge axioms to handle links between these various concepts. We have conjectured the completeness and soundness of our logic. Moreover we have highlighted formal links existing between proper group belief (*i.e.* our group belief operator) and a reductionist group belief formalized here with mutual belief.

We also have seen that strong links appear between group belief and dialogue. In particular group belief is produced thanks to a deliberation process by means of a dialog between group members. This link will be deeply explored in the following part.

Following last remarks of the above section about Tuomela's group belief,

we will in the next chapter introduce the notion of institutions. Indeed when we consider a group of agents and their social relations there is always an (explicit or implicit) institution to manage this relation. We will extend the logic of group belief by introducing the notion of institution. The individual and collective agents will thus have a belief depending on a particular institution or context. This belief depends on a particular context, and it can be argued that it is a kind of acceptance rather than one of belief. This will be explored in the next chapter. We will also see how we can anchor the institution in mental attitudes of the agents.

Chapter 4

An extension: logic of acceptance

To complete our overview of social and collective doxastic states, we present in this chapter the logic $\mathcal{AL}$ (*Acceptance Logic*) in which the acceptance of a proposition by the agents *qua* group members (*i.e.* group acceptance) is introduced. Such propositions are true w.r.t. an institutional context and correspond to facts that are established in an attitude-dependent way (*i.e.* normative and institutional facts). As an application we show how this logic can provide a logical framework for the specification of *autonomous* Multi-Agent Systems (MAS). A MAS is autonomous in so far as it is capable of binding ('nomos') itself ('auto') independently of any external normative constraint specified by a designer. In particular, a MAS is autonomous if it is able to maintain its social institutions (*i.e.* rule-governed social practices) only by way of the agents' attitudes and actions. Finally, we contend that the present approach paves the way for a foundation of legal institutions, for studying the interaction between social and legal institutions and, eventually, for understanding and modeling institutional change.

4.1 Acceptance qua group member

Although in this chapter the notion of acceptance *qua* group member is a primitive (*i.e.* it is not defined in more basic mental attitudes), some conceptual clarification is needed because of the crucial role it plays in the sequel. Whereas beliefs have been studied for decades (Hintikka, 1962) as representative of doxastic mental states, 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 (Velleman, 2000), an acceptance is the output of "a decision to treat p as true in one's utterances and actions" (Hakli, 2006) without being necessarily connected to the actual truth of the proposition. In

order to better specify this distinction, and to recall what has been presented in Section 2.4.3, 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, acceptance 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 chapter we are particularly interested in the last feature, namely the fact that acceptances can be 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 a different one. Although, usually, this aspect of the acceptance state is studied in private contexts (*e.g.* when an agent, in order not to take too many risks, accepts that the total cost of her house restructuring will be beyond her reasonable expectations; see (Bratman, 1992)), we will explore the role of this attitude in institutional contexts and highlight its crucial role in explaining the maintenance of social institutions. Institutional contexts are 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 *explicit* mental states (the acceptances) that can be formally captured. In the context of Clue, for instance, an agent accepts that something has happened (see Example 4.3.3) *qua* player of Clue. The state of acceptance *qua* group member in an institutional context is the kind of acceptance one is committed to when one is "functioning as a group member" (Tuomela, 2007). Although a full analysis of this notion is out of the scope of this dissertation, it is important to stress that we consider this attitude as one that is held by an agent. Nevertheless, there are specific consequences deriving from the agent's functioning as a group member: *e.g.* the acceptance of a proposition *qua* group member is always a public fact (see Section 4.4.1).

In this chapter, we aim to provide a logical framework for the specification of autonomous MASs, that is, MASs whose agents are capable of creating and maintaining their institutions by themselves (Section 4.3). The focus of this contribution is on modeling social or informal institutions, rather than legal ones. Social institutions are the basic structures of a society on top of which more complex legal ones are constructed. By social or informal institutions, we refer to *rule-governed social practices* in which no member with 'special' powers is introduced.[1] More specifically, we will use the notion of an agent's acceptance of a proposition *qua* group member in a given institutional context presented in Section 4.1, and we will study its interaction with different notions such that of common belief and private belief (Section 4.4). On the basis of these attitudes *qua* group members, we will specify how a group can create

[1] It is in fact proper to legal institutions to have specialized agents empowered to change the institution itself on behalf of everybody else (see Section 4.5.1).

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 4.5). We will compare our proposal with related logical works on the issues of collective belief and institutions (Section 4.6). In conclusion we will identify directions for future work on the basis (Section 4.8). Anchoring institutions, and their facts, in agents' minds is just the first step towards a more complete characterization of the "internal aspect" of normative systems and towards the vision of autonomous MASs.

4.2 Institutions

Autonomous agents that interact with each other (and with human beings) pose at least two general problems: they should be able to achieve some level of coordination in order to accomplish their distributed tasks and, notwithstanding their autonomy and self-interest, they should be somehow influenceable towards the fulfillment of some collective goal. One possible way to tackle these problems is to devise artificial *institutions* (Noriega and Sierra, 2002; Dignum and Dignum, 2001). Following the classical work of Douglass North artificial institutions are usually conceived as human-like: "the rules of the game in a society or the humanly devised constraints that structure agents' interaction" (North, 1990, p. 3). With this model in mind, AI practitioners have interpreted their task as that of advancing logical or computational frameworks to represent institutions, while leaving to the agents' autonomy the decision whether to comply or not with the specified rules (Ågotnes et al., 2007; Conte, Castelfranchi, and Dignum, 1999). This approach, however, has at least three strong limitations. First of all, the institutions are not only constraints but also 'enablements' (Searle, 1995): new possibilities of actions (*i.e.* institutional actions like paying, marrying, promising *etc.*) are possible when an institution is in place. Secondly, artificial institutions are usually inspired by human legal institutions which, however, are only a small part of the institutionalized human interactions. Moreover, to work effectively, legal institutions should interact with informal ones.[2] Finally, and more importantly, institutions should be constructed by the agents themselves and not imposed from the outside.

More precisely, while it is widely shared that, in order to face complex and dynamical problems, the individual agents must be autonomous, less emphasis is devoted to the fact that the multi-agent systems (MAS) 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 is the vision of an artificial society that is able to create, maintain, and eventually change its own institutions by itself, without

[2]Following (North, 1990), we consider informal such institutions as social norms and social practices (such as promise). In his seminal book North explicitly states the relevance of this informal layer but this component is still widely neglected in the MAS literature. On the importance of informal normative relations to enforce social order see also (Castelfranchi, 2003).

the intervention of the external designer in this process.

This challenge is also strongly tied to the new trend of designing self-organizing MASs but, in contrast to many efforts in the area, we are after a notion of self-organization that is amenable for, and can make profit of, more complex cognitive agents (*i.e.* BDI-like; see (Conte and Castelfranchi, 1995) for the general approach). In fact, quoting North again (Mantzavinos, North, and Shariq, 2004, 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)."

4.3 The logic

4.3.1 Syntax

The syntactic primitives of our logic $\mathcal{AL}$ (*Acceptance Logic*) are the following:

- a finite set of $n > 0$ agents $AGT = \{1, 2, ..., n\}$;
- a set of atomic formulas $ATM = \{p, q, ...\}$;
- a finite set of labels denoting institutional contexts $INST = \{x, y, ...\}$;
- a symbol λ denoting the private context.

For notational convenience we note $2^{AGT\star} = 2^{AGT} \setminus \{\emptyset\}$ the set of all non empty subsets of agents, $\Delta_1 = \left\{C{:}x | C \in 2^{AGT\star}, x \in INST\right\}$ the set of all couples of non empty subsets of agents and institutional contexts, $\Delta_2 = \{i{:}\lambda | i \in AGT\}$ the set of all couples of single agents and private context, and $i{:}x$ for $\{i\}{:}x$. Finally, $\Delta = \Delta_1 \cup \Delta_2$.

The language $\mathcal{L}_{\mathcal{AL}}$ is defined as the smallest superset of ATM such that: if $\varphi, \psi \in \mathcal{L}_{\mathcal{AL}}$, $i \in AGT$ and $C{:}x \in \Delta$ then $\neg\varphi$, $\varphi \vee \psi$ and $[C{:}x]\,\varphi \in \mathcal{L}_{\mathcal{AL}}$. The classical boolean connectives $\wedge$, $\rightarrow$, $\leftrightarrow$, $\top$ (tautology) and $\bot$ (contradiction) are defined from $\vee$ and $\neg$ in the usual manner.

Formula $[C{:}x]\,\varphi$ has to be read "the agents in C accept that φ while functioning as group members in the institutional context x" or "the group C accepts that φ holds as follower of the institution x".

Example. *$[C{:}Greenpeace]\,protectEarth$ 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 $[i{:}Catholic]\,PopeInfallibility$ is read "the agent i accepts that the Pope is infallible while functioning as a Catholic in the context of the Catholic Church".*

For $C{:}x \in \Delta_1$: $[C{:}x]\,\bot$ has to be read "agents in C are not functioning as group members in the institutional context x" because we assume that functioning as a group member is, at least in this minimal sense, a rational activity;

conversely, $\neg\,[C{:}x]\,\bot$ has to be read "agents in C are functioning as group members in the institutional context x"; $\neg\,[C{:}x]\,\bot \wedge [C{:}x]\,\varphi$ stands for "agents in C are functioning as group members in the context x and they accept that φ while functioning as group members" or simply "agents in C accept that φ *qua* group members in the institutional context x" which, for us, is tantamount to "The group C accepts that φ in the institutional context x" (*i.e.* group acceptance). Similarly, the formula $\neg\,[C{:}x]\,\varphi$ has to be read "agents in C are functioning as group members in the institutional context x and they do not accept that φ while functioning as group members in x" or simply "agents in C do not accept that φ *qua* group members in x" (*i.e.* "The group C does not accept that φ in the institutional context x").

EXAMPLE. $\neg\,[\{i,j\}{:}\mathit{Europe}]\,\bot \wedge [\{i,j\}{:}\mathit{Europe}]\,\mathit{EuroMeansOfExchange}$ *stands for "i and j accept* qua *Europeans that the Euro is the official means of exchange in the context of Europe", whereas* $\neg\,[\{i,j\}{:}\mathit{Europe}]\,\mathit{DollarMeansOfExchange}$ *stands for "i and j* qua *Europeans do not accept that dollar is the official means of exchange".*

Modal operators of the form $[i{:}\lambda]$ correspond to standard doxastic operators.[3] Hence a formula $[i{:}\lambda]\,\varphi$ has to be read "agent i believes that φ".

4.3.2 Semantics

We use a standard possible worlds semantics and a model is a triple $\mathcal{M} = \langle W, \mathscr{G}, \mathscr{V}\rangle$ where:

- W is a set of possible worlds;
- $\mathscr{G} : \Delta \longrightarrow (W \longrightarrow 2^W)$ associates each $C{:}x \in \Delta$ and possible world w with the set $\mathscr{G}_{C:x}(w)$ of possible worlds accepted by the group C in w, where agents in C are functioning as group members in the institutional context x;
- $\mathscr{V}\colon W \longrightarrow 2^{ATM}$ is a truth assignment which associates each world w with the set $\mathscr{V}(w)$ of atomic propositions true at w.

The rules defining the truth conditions of formulas of our logic are inductively 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 $\mathcal{M}, w \models \psi$;
- $\mathcal{M}, w \models [C{:}x]\,\varphi$ iff for all $w' \in W$, if $w' \in \mathscr{G}_{C:x}(w)$ then $\mathcal{M}, w' \models \varphi$.

[3] In this chapter, for the sake of uniformity we prefer to adopt this non-standard notation for doxastic operators instead of $\mathit{Bel}_i\,\varphi$.

4.3.3 Axiomatization

The axiom system of $\mathcal{L}_{\mathcal{AL}}$ is made of all tautologies of propositional calculus, plus the axioms and rules of inference of the basic normal modal logic for every operator $[C{:}x]$ where $C{:}x \in \Delta$. That is, we have all K-theorems for every $C{:}x \in \Delta$.

Moreover, we suppose that given a set of agents C, all $B \subseteq C$ have access to all the facts that are accepted (or that are not accepted) by agents in C while functioning as group members in the institutional context x. In particular, we suppose the following relations between the acceptances of the group members with respect to the institutional contexts: if agents in C (do not) accept that φ while functioning as group members in the institutional context x then for every subset B of C and institutional context y while functioning as group members in the institutional context y, agents in B accept that agents in C (do not) accept that φ while functioning as group members in the institutional context x. Furthermore, we suppose the following relations between the acceptance *qua* group member and individual beliefs: if agents in C (do not) accept that φ while functioning as group members in the institutional context x then, for every agent i in C, we have that i believes that agents in C (do not) accept that φ while functioning as group members in the institutional context x. Finally we suppose standard properties of introspection for beliefs: if agent i believes that φ then he believes that he believes that φ; if agent i does not believe that φ then he believes that he does not believe that φ. Such properties are captured by the following two axiom schemas. For every $C{:}x$, $B{:}y \in \Delta$, if $B \subseteq C$ then:

$$[C{:}x]\,\varphi \rightarrow [B{:}y]\,[C{:}x]\,\varphi \qquad \mathbf{4}_{[C:x],[B:y]}$$

$$\neg\,[C{:}x]\,\varphi \rightarrow [B{:}y]\,\neg\,[C{:}x]\,\varphi \qquad \mathbf{5}_{[C:x],[B:y]}$$

Axioms $\mathbf{4}_{[C:x],[B:y]}$ and $\mathbf{5}_{[C:x],[B:y]}$ together correspond to the following semantic property of Kripke models. For every $w \in W$ and $C{:}x$, $B{:}y \in \Delta$, if $B \subseteq C$ then:

$$\text{if } w' \in \mathscr{G}_{B:y}(w) \text{ then } \mathscr{G}_{C:x}(w') = \mathscr{G}_{C:x}(w) \qquad \mathbf{S1}$$

We also suppose that if agents in C accept that φ *qua* group members in the institutional context x then, for every subset B of C, it holds that agents in B accept φ *qua* group members in the institutional context x. This means that things accepted by the agents in a set C (*qua* group members) with respect to a certain institutional context x are also accepted by agents in all C's subsets with respect to the same context x. Formally, for every $C{:}x$, $B{:}x \in \Delta$, if $B \subseteq C$ then:

$$\neg\,[C{:}x] \bot \wedge [C{:}x]\,\varphi \rightarrow \neg\,[B{:}x] \bot \wedge [B{:}x]\,\varphi \qquad \mathbf{Inc}_{[C:x],[B:x]}$$

EXAMPLE. *Imagine three agents i, j, k that,*qua *players accept, in the context of Clue, that someone called Mrs. Red, has been killed:* $\neg\,[\{i,j,k\}{:}\mathit{Clue}] \bot \wedge [\{i,j,k\}{:}\mathit{Clue}]\,\mathit{killedMrsRed}$. *This implies that also the two agents i, j* qua *Clue players accept that someone called Mrs. Red has been killed in that context:*
$\neg\,[\{i,j\}{:}\mathit{Clue}] \bot \wedge [\{i,j\}{:}\mathit{Clue}]\,\mathit{killedMrsRed}$.

Axiom $\mathbf{Inc}_{[C:x],[B:x]}$ has the following semantic characterization. For every $w \in W$, $C{:}x$, $B{:}x \in \Delta$, if $B \subseteq C$ then:

$$\begin{array}{c} \text{if } \mathscr{G}_{C:x}(w) \neq \emptyset \text{ then } \mathscr{G}_{B:x}(w) \neq \emptyset \\ \text{and } \mathscr{G}_{B:x}(w) \subseteq \mathscr{G}_{C:x}(w) \end{array} \qquad \textbf{S2}$$

As far as operators of type $[i{:}\lambda]$ for beliefs are concerned, we suppose that an agent cannot believe contradictions. Formally, for every $i{:}\lambda \in \Delta_2$:

$$\neg([i{:}\lambda]\,\varphi \wedge [i{:}\lambda]\,\neg\varphi) \qquad \mathbf{D}_{[i:\lambda]}$$

which corresponds to the following standard property of seriality. For every $w \in W$ and $i{:}\lambda \in \Delta_2$ we have:

$$\mathscr{G}_{i:\lambda}(w) \neq \emptyset \qquad \textbf{S3}$$

Thus, every doxastic operator $[i{:}\lambda]$ is *KD45*. (Indeed, besides satisfying Axiom D, it also satisfies Axioms 4 and 5 as particular instances of Axioms $\mathbf{4}_{[C:x],[B:y]}$ and $\mathbf{5}_{[C:x],[B:y]}$ where $C = B = \{i\}$ and $x = y = \lambda$.)

We call $\mathcal{AL}$ (Acceptance Logic) the logic axiomatized by the four principles $\mathbf{4}_{[C:x],[B:y]}$, $\mathbf{5}_{[C:x],[B:y]}$, $\mathbf{Inc}_{[C:x],[B:x]}$, $\mathbf{D}_{[i:\lambda]}$ and we write $\vdash_{\mathcal{AL}} \varphi$ iff formula φ is a theorem of $\mathcal{AL}$. Moreover, let $\mathcal{M}$ be a model such that $\mathcal{M} = \langle W, \mathscr{G}, \mathscr{V} \rangle$ as defined in Section 4.3.2 and satisfying the semantic constraints **S1**–**S3** given above. We write $\models_{\mathcal{AL}} \varphi$ iff formula φ is *valid* in all $\mathcal{AL}$ models, *i.e.* $\mathcal{M}, w \models \varphi$ for every $\mathcal{AL}$ model $\mathcal{M}$ and world w in $\mathcal{M}$. Finally, we say that a formula φ is *satisfiable* if there exists an $\mathcal{AL}$ model $\mathcal{M}$ and a world w in $\mathcal{M}$ such that $\mathcal{M}, w \models \varphi$.

4.4 Group acceptance properties

4.4.1 The public nature of group acceptance

In Section 4.3.1, we have analyzed the notion of group acceptance as the set of the acceptances of all the agents in the group while functioning as group members. This notion of acceptance *qua* group member however must not be confused with (nor reduced to) that of a private mental attitude. On the contrary we claim that group acceptances are always public so much that it is part of the concept of functioning as a group member that all the agents commonly believe that one is functioning in this way. In the literature, an operator to express common belief is given (see for instance (Fagin et al., 1995)). The notion of common belief can be built on the concept of individual belief and on a particular kind of distributed belief of the form "every agent in C believes that φ". The former concept is expressed in our logic by operators of type $[i{:}\lambda]$. The latter concept is formally expressed by operators of type E_C where a formula $E_C\varphi$ is defined as follows:

$$E_C\varphi \stackrel{def}{=} \bigwedge_{i \in C} [i{:}\lambda]\,\varphi$$

Given a set of agents $C \subseteq AGT$, formula $CB_C\varphi$ is meant to stand for "there is common 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 everyone in C believes that φ, and so on". If $E_C^1\varphi$ denotes $E_C\varphi$ and $E_C^k\varphi$ denotes $E_C(E_C^{k-1}\varphi)$, we can define $CB_C\varphi$ as follows[4]:

$$CB_C\varphi \stackrel{def}{=} \bigwedge_{k>0} E_C^k\varphi$$

With the aim of making the public nature of group acceptance explicit, the following theorem highlights the relationship between our notion of group acceptance (*i.e.* acceptance by each of the agents *qua* group members) and the concept of common belief.

THEOREM. *For any $C{:}x \in \Delta$:*

$$\vdash_{\mathcal{AL}} \quad [C{:}x]\,\varphi \leftrightarrow CB_C\,[C{:}x]\,\varphi \tag{4.1}$$

PROOF. *Direction $\rightarrow$ can be established by proving that $\forall k > 0$, $[C{:}x]\,\varphi \rightarrow E_C^k\,[C{:}x]\,\varphi$ by induction on k:*

- $[C{:}x]\,\varphi \rightarrow E_C\,[C{:}x]\,\varphi$ *(case $k = 1$)*
- *From $[C{:}x]\,\varphi \rightarrow E_C^k\,[C{:}x]\,\varphi$ infer* $[C{:}x]\,\varphi \rightarrow E_C^{k+1}\,[C{:}x]\,\varphi$ *(inductive case)*

To prove the case $k = 1$, we just apply Axiom $\mathbf{4}_{[C:x],[B:y]}$ with $B{:}y = i{:}\lambda$ for each $i \in C$, which implies that $[C{:}x] \rightarrow \bigwedge_{i\in C}\,[i{:}\lambda]\,[C{:}x]\,\varphi$. The latter is the case $k = 1$ by definition of E_C.

Let us prove the inductive case. We suppose that $[C{:}x]\,\varphi \rightarrow E_C^k\,[C{:}x]\,\varphi$. By rule of necessitation on every $[i{:}\lambda]$, we infer $\bigwedge_{i\in C}\,[i{:}\lambda]\,([C{:}x]\,\varphi \rightarrow E_C^k\,[C{:}x]\,\varphi)$ which is (by definition of E_C) equivalent to: $E_C([C{:}x]\,\varphi \rightarrow E_C^k\,[C{:}x]\,\varphi)$. Thus from the latter, case $k = 1$ and definition of E_C^{k+1} we can deduce that $[C{:}x]\,\varphi \rightarrow E_C^{k+1}\,[C{:}x]\,\varphi$. This is enough to prove that $[C{:}x]\,\varphi \rightarrow E_C^k\,[C{:}x]\,\varphi$ (for $k > 0$) is a theorem. We can thus infer that $\bigwedge_{k>0}([C{:}x]\,\varphi \rightarrow E_C^k\,[C{:}x]\,\varphi)$ holds. By standard modal principles, $\bigwedge_{k>0}([C{:}x]\,\varphi \rightarrow E_C^k\,[C{:}x]\,\varphi)$ implies $[C{:}x]\,\varphi \rightarrow \bigwedge_{k>0} E_C^k\,[C{:}x]\,\varphi$ which is equivalent to $[C{:}x]\,\varphi \rightarrow CB_C\,[C{:}x]\,\varphi$. We leave to the reader the proof of $\leftarrow$ direction of the theorem. □

According to Theorem 4.1, the agents in C accept that φ while functioning as group members in the institutional context x if and only if there is common belief in C that they accept that φ while functioning as group members in the institutional context x. Hence, accepting a proposition while functioning as a group member is always a *public* fact which is out in the open and that is used by all the members to reason about each other in an institutional context.

[4]Note that this definition is the iterative version of the one given in the above chapter.

4.4.2 Group acceptance and individual beliefs

As far as the relationship between acceptances *qua* group members and individual beliefs is concerned, it has to be noted that $\neg [C{:}x] \bot \wedge [C{:}x] \varphi \wedge \bigwedge_{i \in C} [i{:}\lambda] \neg \varphi$ where $C{:}x \in \Delta_1$ is satisfiable in our logic. This means that the attitudes privately endorsed by the agents and those entertained *qua* group members can diverge: one can privately disbelieve what one accepts while functioning as a group member.

EXAMPLE. *Consider the discursive dilemma as elaborated in (Pettit, 2001) in which a three-member court has to make a judgment on whether a defendant is liable for a breach of contract. If one assumes that the group accepts the majority rule to decide on the issue, it might happen that each judge privately believes that the group ought to accept a certain conclusion (*e.g. *that the defendant is liable), while each is forced to accept the opposite* qua *group member (*i.e. qua *judge).*

4.5 Attitude-dependent facts

Normative and institutional facts are a class of facts that are typical of institutional contexts (Searle, 1995). Such facts have the peculiar feature of being dependent on the agents' attitudes in a way that we are now in the position to specify in detail. More precisely it has been noted that these facts are characterized at least by two features (Lagerspetz, 1995; Searle, 1995; Tuomela, 2002).

- Performativity: an attitude of certain type 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.

EXAMPLE. *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).*

In order to represent in $\mathcal{AL}$ these kinds of facts, we need first to define the concept of truth with respect to an institutional context in a way that respects these two principles.

4.5.1 Truth in an institutional context

We formalize the notion of truth w.r.t. to a certain institutional context with the operator $[x]$. A formula $[x]\varphi$ is read "within the institutional context x, it

is the case that φ". Here we suppose that "within the institutional context x it is the case that φ" if and only if "for every set of agents C, the agents in C accept that φ while functioning as group members in the institutional context x". Formally, for $x \in INST$:

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

It is straightforward to prove that $[x]$ are normal modal operators. Given the previous analysis, a fact is true w.r.t. an institutional context if and only if such fact is accepted by all the agents while they function as group members (hence the performativity and the reflexivity principles are maintained). Moreover, following Theorem 4.1, this group acceptance is the object of a common belief.

At this point, it might be objected that there are facts which are true in an institutional context but only "special" group members in the institution are aware of them. For instance, there are laws in every country which are known only by the specialists of the domain (lawyers, judges, members of the parliament, *etc.*). Are these facts accepted by the institution (here the country) notwithstanding that many group members are not aware of them?

In order to resist to this objection recall that, at this stage, our model applies to the basic informal institutions of a society. Relative to this restriction, the proposed assumption is justified because, w.r.t. these institutions, there is no other special institutional contexts in which the agents have the power to create and eliminate institutional facts characterizing the institution itself (*i.e.* nobody has the power to change the rules for promising). It is in fact peculiar of legal (formal) institutions to create such a specialized *meta*-context in which the agents have special powers to interpret and modify the institution itself. Given the aims of this paper, we leave this special case for future work.

Finally, the following abbreviation is defined:

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

which stands for "φ is universally accepted as true". It is straightforward to prove that $[Univ]$ is a normal modal operator.

4.5.2 Contextual conditionals

From the concept of truth with respect to an institutional context a notion of *contextual conditional* can be defined. A contextual conditional is a material implication of the form $\varphi \rightarrow \psi$ in the scope of an operator $[x]$. A contextual conditional is a local one, that is, a conditional that is not universally valid while it is accepted by the group members in a specific institutional context. More precisely, we exclude the situation in which $[Univ]\,(\varphi \rightarrow \psi)$ is true.

Example.

Let us consider the institutional context of gestural language in Europa. There exists a contextual conditional in this language according to which, the

nodding gesture "counts as" an endorsement of what the speaker is suggesting. This conditional is formally expressed by the construction $[gesture]\,(nodding \rightarrow yes)$. *It is clear that this kind of conditional is not universally valid (*e.g. *in a different cultural context the same gesture may express exactly the opposite fact). Thus,* $\neg\,[Univ]\,(nodding \rightarrow yes)$ *holds.*

More generally, for every $x \in INST$ we define the following abbreviation:

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

$\varphi \overset{x}{\rhd} \psi$ stands for "in the institutional context x, if φ then ψ". Although the presentation and the discussion of all relevant properties of our construction $\varphi \overset{x}{\rhd} \psi$ is out of the scope of this chapter, it is interesting to note that $\varphi \overset{x}{\rhd} \psi$ satisfies some intuitive properties of counts-as conditionals as identified in (Jones and Sergot, 1996).

THEOREM. *For every* $x \in INST$*:*

$$\textit{From } \vdash_{\mathcal{AL}} (\varphi_2 \leftrightarrow \varphi_3) \textit{ infer } \vdash_{\mathcal{AL}} (\varphi_1 \overset{x}{\rhd} \varphi_2 \leftrightarrow \varphi_1 \overset{x}{\rhd} \varphi_3) \tag{4.2}$$

$$\textit{From } \vdash_{\mathcal{AL}} (\varphi_1 \leftrightarrow \varphi_3) \textit{ infer } \vdash_{\mathcal{AL}} (\varphi_1 \overset{x}{\rhd} \varphi_2 \leftrightarrow \varphi_3 \overset{x}{\rhd} \varphi_2) \tag{4.3}$$

$$\vdash_{\mathcal{AL}} (\varphi_1 \overset{x}{\rhd} \varphi_2 \wedge \varphi_1 \overset{x}{\rhd} \varphi_3) \rightarrow (\varphi_1 \overset{x}{\rhd} (\varphi_2 \wedge \varphi_3)) \tag{4.4}$$

$$\vdash_{\mathcal{AL}} (\varphi_1 \overset{x}{\rhd} \varphi_2 \wedge \varphi_3 \overset{x}{\rhd} \varphi_2) \rightarrow ((\varphi_1 \vee \varphi_3) \overset{x}{\rhd} \varphi_2) \tag{4.5}$$

$$\vdash_{\mathcal{AL}} (\varphi_1 \overset{x}{\rhd} \varphi_2 \wedge (\varphi_1 \wedge \varphi_2) \overset{x}{\rhd} \varphi_3) \rightarrow (\varphi_1 \overset{x}{\rhd} \varphi_3) \tag{4.6}$$

PROOF. *We only provide a proof of Theorem 4.6 as an example. This theorem expresses a property of cumulative transitivity (cut). The other theorems and rules of inference can be proved straightforwardly by definition of* $\varphi \overset{x}{\rhd} \psi$*) and the axioms and rules of inference of* $\mathcal{AL}$. $\varphi_1 \overset{x}{\rhd} \varphi_2 \wedge (\varphi_1 \wedge \varphi_2) \overset{x}{\rhd} \varphi_3$ *implies* $[x]\,(\varphi_1 \rightarrow \varphi_2)$ *and* $[x]\,((\varphi_1 \wedge \varphi_2) \rightarrow \varphi_3)$ *which in turn imply* $[x]\,(\varphi_1 \rightarrow \varphi_3)$ *(by the fact that* $[x]$ *is normal. Moreover,* $\varphi_1 \overset{x}{\rhd} \varphi_2 \wedge (\varphi_1 \wedge \varphi_2) \overset{x}{\rhd} \varphi_3$ *implies* $\neg\,[Univ]\,((\varphi_1 \wedge \varphi_2) \rightarrow \varphi_3)$ *which is equivalent to* $\neg\,[Univ]\,(\neg\varphi_1 \vee \neg\varphi_2 \vee \varphi_3)$. *As* $[Univ]$ *is also a normal modal operator,* $\neg\,[Univ]\,(\neg\varphi_1 \vee \neg\varphi_2 \vee \varphi_3)$ *implies* $\neg\,[Univ]\,(\neg\varphi_1 \vee \varphi_3)$ *(by the fact that* $[Univ]$ *is normal) which in turn is equivalent to* $\neg\,[Univ]\,(\varphi_1 \rightarrow \varphi_3)$. □

Moreover, we can easily show that our concept of contextual conditional does not satisfy reflexivity, transitivity and weakening of the antecedent, that is, the following three formulas are not valid: $\varphi \overset{x}{\rhd} \varphi$, $(\varphi_1 \overset{x}{\rhd} \varphi_2 \wedge \varphi_2 \overset{x}{\rhd} \varphi_3) \rightarrow \varphi_1 \overset{x}{\rhd} \varphi_3$, and $\varphi_1 \overset{x}{\rhd} \varphi_2 \rightarrow (\varphi_1 \wedge \varphi_3) \overset{x}{\rhd} \varphi_2$. As discussed in Section 4.6.2 our notion of contextual conditional is similar to the notion of *proper classificatory rule* given in (Grossi, 2006). [5]

[5] We refer to (Grossi, 2006) for interesting arguments concerning why proper classificatory rules should not necessarily satisfy reflexivity, transitivity and weakening of the antecedent.

4.5.3 Normative facts

While contextual conditionals are useful to understand the notion of institutional facts, they are not sufficient for a more precise characterization. In fact, as noted in (Searle, 1995), institutional facts are always connected to a deontic dimension that up to now is still missing.

In our perspective, a contextual conditional $\varphi \overset{x}{\rhd} \psi$ can be adopted to represent an institutional fact if and only if the term ψ in the contextual conditional is a fact to which a certain number of obligations and permissions are associated within the institutional context x. In this sense, ψ is an institutional fact with respect to the institutional context x.

EXAMPLE. *"Being eighteen years old counts as being of age" is a constitutive rule accepted by a set of agents* qua *citizens in Italy and "being of age" is an institutional fact with respect to this context. Moreover, to such an institutional fact a certain number of permissions and obligations are associated (*e.g. *in Italy if you are of age you have the permission to vote and the obligation to fulfill the military duties). In this sense the constitutive rule "being eighteen years old counts as being of age" connects the institutional fact "being of age" with the brute fact "being eighteen years old" which is a fact intrinsically connected to certain normative facts.*

In order to capture this core feature, our logic $\mathcal{AL}$ can be appropriately extended by introducing a *violation* atom V as in Anderson's reduction of deontic logic to alethic logic (Anderson, 1958) and in dynamic deontic logic (Meyer, 1988). By means of this new formal construct we can specify normative facts (*i.e.* what it is obligatory and permitted) in a way that respects their being also a kind of attitude-dependent fact holding relative to certain attitudes and in a specific institutional context. As far as obligations are concerned, we say that "φ is something obligatory within the institutional context x" (noted $O(\varphi, x)$) if and only if "$\neg\varphi \rightarrow V$ is a contextual conditional in the institutional context x" or, more specifically, "$\neg\varphi$ counts as a violation within the institutional context x". Formally:

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

As far as permission is concerned we say that "φ is something permitted within the institutional context x" (noted $P(\varphi, x)$) if and only if $\neg\varphi$ is not obligatory within the institutional context x. Formally:

$$P(\varphi, x) \overset{def}{=} \neg O(\neg\varphi, x)$$

Formulas of type $O(\varphi, x)$ and $P(\varphi, x)$ can be conceived as particular instances of so-called *regulative rules*, that is, rules which specify the ideal behavior of agents in terms of permissions, obligations, and prohibitions. We refer to these rules as normative facts.[6]

[6] The distinction between *regulative rule* and *constitutive rule* has been emphasized by Searle (Searle, 1995) and then modelled in logic by several authors. For an example see (Boella and van der Torre, 2004).

EXAMPLE. *The formula $O(driveCar \rightarrow \neg RightSide, UK)$ is a normative fact in the UK within whose context it is obligatory to drive on the left side of the street (*i.e. *"driving a car on the right side of the street counts as violation in UK").*

Again, it is important to stress the fact that normative facts, by being represented with a contextual conditional are attitude-dependent facts and are intrinsically connected with the acceptance of all the agents *qua* group members in a specific institutional context.

4.5.4 Institutional facts and constitutive rules

We are now in the position to formalize what an institutional fact is. Let $2^{\mathcal{L}_{\mathcal{AL}}\star} = 2^{\mathcal{L}_{\mathcal{AL}}} \setminus \{\emptyset\}$ be the set of non empty subsets of $\mathcal{L}_{\mathcal{AL}}$. From the previous construction $\varphi \overset{x}{\rhd} \psi$ it is straightforward to come up with a formal characterization of nonempty sets of formula. Formally, for every $x \in INST$ and Σ_O, $\Sigma_P \in 2^{\mathcal{L}_{\mathcal{AL}}\star}$:

$$InstFact_x^{\Sigma_O,\Sigma_P}(\varphi) \overset{def}{=} \bigwedge_{\sigma\in\Sigma_O} O(\varphi \rightarrow \sigma, x) \wedge \bigwedge_{\sigma'\in\Sigma_P} P(\varphi \wedge \sigma', x)$$

$InstFact_x^{\Sigma_O,\Sigma_P}(\varphi)$ stands for "φ is an institutional fact within the institutional context x characterized by the set of obligations Σ_O and the set of permissions Σ_P".

EXAMPLE. *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".*[7]

From the concept of institutional fact we can also formalize the concept of constitutive rule. To this aim, we must make explicit the fact that the term ψ in $\varphi \overset{x}{\rhd} \psi$ is an institutional fact to which a set of obligations and a set of permissions are associated. Formally, for every $x \in INST$ and Σ_O, $\Sigma_P \in 2^{\mathcal{L}_{\mathcal{AL}}\star}$:

$$ConstRule_x^{\Sigma_O,\Sigma_P}(\varphi,\psi) \overset{def}{=} \varphi \overset{x}{\rhd} \psi \wedge InstFact_x^{\Sigma_O,\Sigma_P}(\psi)$$

$ConstRule_x^{\Sigma_O,\Sigma_P}(\varphi,\psi)$ stands for "φ counts as ψ is a constitutive rule of institution x where ψ is an institutional fact within the institutional context x characterized by the set of obligations Σ_O and the set of permissions Σ_P".

EXAMPLE. *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*

[7] A more precise formulation of this example needs a representation of the right relation which is, however, beyond the scope of this article. See (Makinson, 1986) for more details.

military duties". In this sense $ConstRule_{Italy}^{\{military\},\{vote\}}(eighteen, toBeOfAge)$ *is a specific kind of contextual conditional in which the connection between the institutional fact toBeOfAge and the brute fact eighteen is established. A number of normative facts consisting in obligations and permissions pertain to the institutional fact toBeOfAge, namely* $O(toBeOfAge \rightarrow military, Italy)$ *and* $P(toBeOfAge \wedge vote, Italy)$.

4.6 Related works

4.6.1 Link between $\mathcal{AL}$ and the G logic

We now compare group acceptance presented in the current chapter with group belief described in the previous one. We recall that: $G_C\,\varphi$ means that "it is collectively believed by group C that φ is true". When C is reduced to a singleton $\{i\}$, $G_{\{i\}}$ is identified with the belief *à la* Hintikka (Hintikka, 1962). In this view, group belief is rational ($\mathbf{D}_{G_C}$), public for every subgroup ($\mathbf{SR}_+$ and $\mathbf{SR}_-$) and it has been formed by the joint acceptance of all members (**WR** and **CG**). We recall the axiomatics of Chapter 3:

($\mathbf{D}_{G_C}$) $G_C\,\varphi \rightarrow \neg G_C\,\neg\varphi$
($\mathbf{SR}_+$) $G_C\,\varphi \rightarrow G_{C'}\,G_C\,\varphi, C' \subseteq C$
($\mathbf{SR}_-$) $\neg G_C\,\varphi \rightarrow G_{C'}\,\neg G_C\,\varphi, C' \subseteq C$
(**WR**) $G_C\,\varphi \rightarrow G_C\,G_{C'}\,\varphi$, is $C' \subseteq C$ and φ objective.
(**CG**) $(\bigwedge_{i\in C} G_C\,G_i\,\varphi) \rightarrow G_C\,\varphi$

Notions of group belief and group acceptance seem to be very close. Thus the idea of expressing the G operator in $\mathcal{AL}$ appears intuitive because $\mathcal{AL}$ is more expressive, with the notion of context lacking in the G logic. We show in the sequel that $\mathcal{AL}$ can subsume in some way G's logic.

4.6.1.1 Representing G operator in $\mathcal{AL}$

In Chapter 3, we did not take explicitly into account the context, or more precisely we did not distinguish the group from the context in which the dialogue took place. For example, if we would like to formalize that the group C, as followers of Greenpeace, believes that earth must be protected, we can only write: $G_{Greenpeace}\,protectEarth$ and deduce thank to (**WR**) that $G_{Greenpeace}\,G_C\,protectEarth$, with $C \subseteq Greenpeace$. As showed in this example, we reduce the institution *Greenpeace* to the group of its members. This is inappropriate in this case because the institution Greenpeace exists independently of agents composing it.

Thus the G operator does not take into account various institutional contexts, or in other words it considers (implicitly) only one. Thus formally we have:

$$G_C\,\varphi \equiv [C{:}x_C]\,\varphi$$

where x_C is the only institution whereby C is concerned and where $x_{\{i\}} \equiv \lambda$. The context x_C associated to each group of agents represents the C's internal

institution, *i.e.* the framework of the dialog between members of the group. It is thus closely related to the particular group and unique for this group.

We need both to examine and compare axiomatics. Axioms $\mathbf{4}_{[C:x],[B:y]}$ and $\mathbf{5}_{[C:x],[B:y]}$ are generalizations of the ($\mathbf{SR}_+$) and ($\mathbf{SR}_-$) for contexts x_C and x_B instead of x and y. They represent the public nature of both notions. Axiom $\mathbf{Inc}_{[C:x],[B:x]}$ cannot be expressed in the grounding logic, except under his tautological and uninformative form where $B = C$. An axiom such as: $G_C\,\varphi \rightarrow G_B\,\varphi$, would be too strong because we consider that belief of a subgroup is not related to the supergroup's beliefs (and in particular group belief is totally independent of individual group members' beliefs).

Some axioms lack in $\mathcal{AL}$ to exactly capture the G operators. In particular the axiom $\mathbf{D}_{[i:\lambda]}$ should be generalized to $[C{:}x_C]$ representing that agents in C are *de facto* functioning as group members in the context x_C. As $x_{\{i\}} = \lambda$, axiom ($\mathbf{D}_{[C:x_C]}$) represents that group belief and individual belief have same features, in particular that group C is *de facto* follower of the context x_C.

$\mathbf{D}_{[C:x_C]}$ $\neg\,[C{:}x_C]\,\bot$, for $C \in 2^{AGT} \setminus \emptyset$

Moreover axioms (**WR**) and (**CG**) express that a group belief is established by a consensus of expressed opinion. They do not have a counterpart in the $\mathcal{AL}$ logic, because we are only concerned here by properties of acceptance (not by its formation). (**WR**) and (**CG**) could be translated directly.

WR $[C{:}x_C]\,\varphi \rightarrow [C{:}x_C]\,[C' : x_{C'}]\,\varphi$, if $C' \subseteq C$ and φ is objective.

CG $(\bigwedge_{i\in C}\,[C{:}x_C]\,[i{:}\lambda]\,\varphi) \rightarrow [C{:}x_C]\,\varphi$

(**CG**) would express that a formula φ is accepted by a group C in the particular context x_C if C has accepted that every C's members believe personally that φ (which does not imply that they actually believe it). These three additional axioms are due to the features of the particular context x_C: they represent the strong link existing between x_C and C. We can note that Theorem 4.1 is also a theorem of the grounding logic. In the sequel, we explore interactions between G_C defined as $[C{:}x_C]$ and general acceptance $[C{:}x]$, which produces mixed theorems.

4.6.1.2 Extension of the integration

As $\neg\,[C{:}x]\,\bot \rightarrow \neg\,[B{:}x]\,\bot$ (with $B \subseteq C$) is a theorem of $\mathcal{AL}$ (the proof can be easily built from $\mathbf{Inc}_{[C:x],[B:x]}$[8]), we have: $\neg\,[C{:}x_C]\,\bot \rightarrow \neg\,[B{:}x_C]\,\bot$, with

[8]

1. $\vdash \top$
2. $\vdash [C{:}x]\,\top$, by $\mathrm{RN}_{[C:x]}$
3. $\vdash \neg\,[C{:}x]\,\bot \wedge [C{:}x]\,\varphi \rightarrow \neg\,[B{:}x]\,\bot \wedge [B{:}x]\,\varphi$, by ($\mathbf{Inc}_{[C:x],[B:x]}$)
4. $\vdash \neg\,[C{:}x]\,\bot \wedge [C{:}x]\,\top \rightarrow \neg\,[B{:}x]\,\bot \wedge [B{:}x]\,\varphi$, from 3. by LP
5. $\vdash [C{:}x]\,\top \rightarrow (\neg\,[C{:}x]\,\bot \rightarrow \neg\,[B{:}x]\,\bot)$, by 4. and LP
6. $\vdash \neg\,[C{:}x]\,\bot \rightarrow \neg\,[B{:}x]\,\bot$, by 2., 5. and Modus Ponens

$B \subseteq C$, which means that all agents in subsets of C are also functioning as group members in the context x_C.

Moreover, as $\neg [C{:}x_C] \bot$ is valid, Axiom $\mathbf{Inc}_{[C:x],[B:x]}$ is reduced to the following theorem: $[C{:}x_C] \varphi \rightarrow [B{:}x_C] \varphi$, with $B \subseteq C$. Thus if φ is a belief of the group C, every subgroup accepts it in the context of x_C; there is a group acceptance on what is collectively believed. For example, if it is collectively believed by the activists that the aim of Greenpeace is to protect the Earth then in the context of Greenpeace every subgroup must accept it. This does not imply anything about subgroup and individual beliefs.

From the previous theorem, we can also prove that:

$$[C{:}x_C] \varphi \rightarrow [C{:}x_C] [B{:}x_C] \varphi, \text{ with } B \subseteq C$$

PROOF.

1. $\vdash [C{:}x_C] \varphi \rightarrow [B{:}x_C] \varphi$, *by the above theorem*
2. $\vdash [C{:}x_C] [C{:}x_C] \varphi \rightarrow [C{:}x_C] [B{:}x_C] \varphi$, *by 1. and* ($\boldsymbol{RN}_{[C:x_C]}$)
3. $\vdash [C{:}x_C] \varphi \rightarrow [C{:}x_C] [C{:}x_C] \varphi$, *by* ($\boldsymbol{4}_{[C:x],[B:y]}$)
4. $\vdash [C{:}x_C] \varphi \rightarrow [C{:}x_C] [B{:}x_C] \varphi$, *by 2., 3. and LP*

□

This theorem extends the previous one: if φ is collectively believed, every subgroup accepts φ in the context x_C (by the former theorem), but this acceptance is also collectively believed. This theorem is in fact quite close to axiom (**WR**) in the grounding logic.

4.6.2 Related works on normative systems

Because of interesting formal similarities, we will just compare $\mathcal{AL}$ with (Grossi, 2006) in which a modal logic for the formalization of count-as assertions and the specification of normative systems has been proposed. This logic is based on a set of modal operators $[x]^*$ where the index x is in a set of indexes C_0.[9] An index x is supposed to denote a certain institutional context (or normative system). Operators $[x]^*$ are similar to our operators $[x]$ defined in Section 4.5.1. A formula $[x]^* \varphi$ approximately stands for "in the institutional context/normative system x it is the case that φ". An operator $[u]^*$ is also used for denoting facts which universally hold. The set $C = C_0 \cup \{u\}$ is given by adding index u to the set of indexes C_0. Differently from our logic where the contextual operator $[x]$ is built on the notion of group acceptance, in Grossi's logic the contextual operator $[x]^*$ is given as a primitive operator. Operators $[x]^*$ and $[u]^*$ are exploited in Grossi's logic to define contextual conditionals called *proper classificatory rules* noted by $\varphi \Rightarrow_i^{cl+} \psi$ which is an abbreviation of $[x]^* (\varphi \rightarrow \psi) \wedge \neg [u]^* (\varphi \rightarrow \psi)$

[9] Here we use the notation $[x]^*$ in order to distinguish Grossi's operators from our operators $[x]$.

and is meant to stand for "φ counts as ψ in the normative system x". The construction $\varphi \Rightarrow_i^{cl+} \psi$ is similar to our $\varphi \overset{x}{\rhd} \psi$.[10] Operator $[u]^*$ is $S5$ and the logic is supposed to satisfy the following additional principles. For any $x, y \in C$:

1. $[x]^* \varphi \rightarrow [y]^* [x]^* \varphi$
2. $\neg [x]^* \varphi \rightarrow [y]^* \neg [x]^* \varphi$
3. $[u]^* \varphi \rightarrow [x]^* \varphi$
4. $[u]^* \varphi \rightarrow \varphi$

According to 1. and 2., truth and falsehood in institutional contexts/normative systems are absolute because they remain invariant even if 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. These two principles are in our view criticizable because they rely on the very counter-intuitive assumption that all facts true in an institutional context are public to all other institutional contexts. But, what does it mean that a fact is known by an institution? Our aim here is to show that such an assumption can be disambiguated in our logical framework. The relevant question is: under what additional assumptions formulas $[x]\varphi \rightarrow [y][x]\varphi$ and $\neg [x]\varphi \rightarrow [y]\neg [x]\varphi$ can be inferred in our logic? On the one hand, it is easy to prove that the principles given in Section 4.3.3 are not sufficient to infer such formulas. Indeed, formulas $[x]\varphi \wedge \neg [y][x]\varphi$ and $\neg [x]\varphi \wedge \neg [y]\neg [x]\varphi$ are satisfiable in $\mathcal{AL}$. On the other hand, it is straightforward to show that: if Axioms $\mathbf{4}_{[C:x],[B:y]}$ and $\mathbf{5}_{[C:x],[B:y]}$ are weakened by supposing that they **also** hold for $B \nsubseteq C$, then formulas $[x]\varphi \rightarrow [y][x]\varphi$ and $\neg [x]\varphi \rightarrow [y]\neg [x]\varphi$ can be inferred. This means that in our logic Grossi's properties can be derived under the assumption that, given two arbitrary sets of agents B and C, agents in B have access to all facts that agents in C accept (do not accept), while functioning as group members in a certain institutional context x. That is, given an arbitrary set of agents C, if agents in C accept that φ while functioning as a group members in the institutional context x then this fact is public in such a way that all other agents outside C accept that agents in C accept that φ while functioning as group members in the institutional context x.

Concerning the principle 4., it says that: if φ universally holds then φ is true. This principle is also criticizable in our opinion. For instance, during the 7th-6th century BC people believed that the earth was flat. But it has never been the case that earth was/is/will be flat.

More generally, if we suppose that: Axioms $\mathbf{4}_{[C:x],[B:y]}$ and $\mathbf{5}_{[C:x],[B:y]}$ studied in Section 4.3.3 are also valid for $B \nsubseteq C$; the T axiom is valid for $[Univ]$ operator (in a similar way of the previous principle 4.); and the following translations of Grossi's operators $[x]^*$ and $[u]^*$ into our logic $\mathcal{AL}$ are given

- $tr([x]^* \varphi) = [x]\varphi$
- $tr([u]^* \varphi) = [Univ]\varphi$,

[10] The author distinguishes *proper classificatory rules* from mere *classificatory rules* and *constitutive rules*. Differently from *classificatory rules*, *proper classificatory rules* are rules which would not hold without the normative system/institution stating them. In (Grossi, Meyer, and Dignum, 2006) a further distinction between *classificatory rules* and *constitutive rules* is given.

we can prove that the translations into $\mathcal{AL}$ of all of Grossi's axioms are $\mathcal{AL}$ theorems.

THEOREM. *Suppose that: i)* $[Univ]\,\varphi \rightarrow \varphi$ *is valid, and that for every* $C{:}x, B{:}y \in \Delta$, *ii)* $[C{:}x]\,\varphi \rightarrow [B{:}y]\,[C{:}x]\,\varphi$ *and iii)* $\neg\,[C{:}x]\,\varphi \rightarrow [B{:}y]\,\neg\,[C{:}x]\,\varphi$ *are valid in* $\mathcal{AL}$. *Then, the following properties can be inferred in* $\mathcal{AL}$*:*

- $[x]\,\varphi \rightarrow [y]\,[x]\,\varphi$
- $\neg\,[x]\,\varphi \rightarrow [y]\,\neg\,[x]\,\varphi$
- $[Univ]\,\varphi \rightarrow [x]\,\varphi$
- $[Univ]$ *satisfies all axioms and rules of inference of the system* $S5$

PROOF. *We only provide a proof of the last item of the theorem. The other items can be proved in a similar way. First of all* $[Univ]$ *is a normal modal operator by definition as a conjunction of normal modal operators* $[x]$. *We have property* $\mathbf{T}_{[Univ]}$ *by Hypothesis i). We only need to prove that* $\mathbf{4}_{[Univ]}$ *and* $\mathbf{5}_{[Univ]}$ *can be inferred from the hypotheses. From Hypothesis ii) we can deduce that* $[C{:}x]\,\varphi \rightarrow \bigwedge_{B:y\in\Delta}[B{:}y]\,[C{:}x]\,\varphi$ *which is equivalent (by definition of* $[Univ]\,\varphi$*) to* $[C{:}x]\,\varphi \rightarrow [Univ]\,[C{:}x]\,\varphi$, *which entails* $\bigwedge_{C:x\in\Delta}[C{:}x]\,\varphi \rightarrow \bigwedge_{C:x\in\Delta}[Univ]\,[C{:}x]\,\varphi$, *which is equivalent to* $[Univ]\,\varphi \rightarrow [Univ]\,[Univ]\,\varphi$ (i.e. $\mathbf{4}_{[Univ]}$). $\mathbf{5}_{[Univ]}$ (i.e. $\neg\,[Univ]\,\varphi \rightarrow [Univ]\,\neg\,[Univ]\,\varphi$) *can be inferred from Hypothesis iii) in a similar way.* □

4.7 An attempt toward formal institutions

In this section we give only starting ideas to extend our $\mathcal{AL}$ logic to take into account formal institutions.

4.7.1 A sophistication: Legislators

For formal institutions where there is some leader who has the power over the legal system, the previous definition of "truth with respect to an institution" must be appropriately redefined after introducing a concept of legislator.

Suppose a function *Leg* is introduced, where *Leg* stands for the legislator(s). This function assigns a group of agents in AGT to every institution x: $Leg : INST \longrightarrow 2^{AGT}$. $Leg(x) = C$ means that agents in C are the legislator of institution x, that is, agents in C are legally responsible over x. It seems reasonable to suppose that the legislator of a certain institution x always share a common view as followers of institution x, otherwise they would not be the legislator(s) of x. This assumption is expressed by the following axiom:

$$\neg\,[Leg(x) : x]\,\bot$$

Given function *Leg* and the previous assumption the notion of truth w.r.t. to an institution can be refined as follows:

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

This means that "according to institution x it is the case that φ" if and only if "the legislator(s) of institution x jointly accept that φ as followers of institution x".

4.7.2 From social roles to institutional powers

In particular to represent institutional powers into our logic, we need to extend it by introducing actions and roles. We could introduce actions in $\mathcal{AL}$ logic in the same manner as for group belief logic (*cf.* Section 3.2.5), by adding a accessibility relation and its converse to the model. We thus have in particular the following operators: $After_\alpha\,\varphi$ meaning "φ holds after every execution of α" and $Happens_\alpha\,\varphi$ meaning "α is happening and φ is true just afterwards".

As in (Grossi et al., 2005), we could introduce in our logic a set $\mathcal{R}$ of social role labels. Elements in $\mathcal{R}$ are noted by $r_1, r_2, ...r_m$. We suppose that at world w in a model M an agent $i \in AGT$ has a certain role $r_j \in \mathcal{R}$ in a certain institution $x \in INST$ if and only if formula $Role(i, r_j, x)$ is true. Formula $Role(i, r_j, x)$ is evaluated according to the following function: $f : \mathcal{R} \times INST \longrightarrow 2^{AGT}$. The truth condition of $Role(i, r_j, x)$ is as follows: $M, w \models Role(i, r_j, x) \Longleftrightarrow i \in f(r_j)$

From the previous definition of social role we can extract a quite general notion of institutional power assigned to social roles. For any $x \in INST$, $r_j \in \mathcal{R}$ and $a \in ACT$:

$$Power(r_j, a, \varphi, x) \stackrel{def}{=} \bigwedge_{i \in AGT} (Role(i, r_j, x) \stackrel{x}{\rhd} After_{i:a}\varphi)$$

$Power(r_j, a, \varphi, x)$ has to be read "according to institution x an agent playing social role r_j has the institutional power of ensuring φ by doing action a". For example, $Power(priest, gesture, married, church)$ stands for "in the church the priest has the power of marrying a couple by performing certain gestures".

Nevertheless, there are two different and equally important views of institutional power. On the one hand, institutional power can be conceived as a capacity that agents playing a certain social role r_j have according to a certain institution x. On the other hand, institutional power can be conceived as an agent i's exercise of a capacity that i has due to the fact that his playing a certain social role r_j in an institution x. The notion of institutional power defined above is a power in the former sense. Now, we want to look at institutional power in the latter sense. To this end, we provide a quite general definition of *exercise of institutional power*. We suppose that an agent i playing a role r_j in institution x exercises his power of ensuring φ by doing action a if and only if according to institution x agent i playing social role r_j has the institutional power of ensuring φ by performing action a and, according to institution x agent

i performs action a by playing role r_j. Formally, for any $x \in INST$, $r_j \in \mathcal{R}$, $i \in AGT$ and $a \in ACT$:

$$ExPower(i, r_j, a, \varphi, x) \stackrel{def}{=} Power(r_j, a, \varphi, x) \wedge Role(i, r_j, x) \wedge [x]\, Happens_{i:a}\, \top$$

For example, $ExPower(i, priest, gesture, married, church)$ stands for "in the church agent i playing the role of priest exercises his power of marrying a couple by performing certain gestures".

The exercise of an institutional power by a certain agent playing a certain role modifies the current asset of the normative system by creating new obligations and permissions and by removing previous obligations and permissions. This dynamic aspect of institutions, although it could be analyzed in our framework, remains out of the scope of this dissertation.

4.8 Conclusion

We have started this chapter by proposing a Group Acceptance Logic, the $\mathcal{AL}$ logic in which the agents' attitudes *qua* group members can be analyzed. We use it to raise the challenge of *autonomy* at the level of MASs, so that they will be able to bind themselves in ways that further the achievement of collective goods in dynamic and uncertain environments as human societies do. Given the properties of a demystified notion of group acceptance in an institutional context, we have provided an analysis of the kind of attitude-dependent facts typical of institutions. In particular, we have introduced a notion of obligation and permission with respect to an institutional context (*i.e.* so-called normative facts). Then, we have defined institutional facts. In our perspective an institutional fact within the institutional context x is a fact to which a number of obligations and permissions are (contextually) associated. Finally, we have formalized the concept of constitutive rule, that is, a rule which is responsible for the connection between an institutional fact and a brute physical fact.

In our view, a constitutive rule is a rule of type "φ counts as ψ in the institutional context x" where ψ denotes an institutional fact within the institutional context x. While such rules are usually defined from the external perspective of a normative system or institution, we have, once again, anchored these rules in the agents' attitudes.

Although the present model is focused on the neglected layer of informal institutions, it still lacks sufficient expressiveness to represent the phenomenon of "institutionalized power" (Jones and Sergot, 1996) which is, of course, crucial also within this kind of institutions. In order to cope with limitation, in future work, we will expand $\mathcal{AL}$ with Propositional Dynamic Logic (PDL) in order to be able to talk about actions within our language. Moreover, a first kind of dynamics will be studied in which agents, *qua* group members in specific institutional contexts, will be able to create new institutional facts. Given the way we have modeled such facts, the agents will update and revise their own deontic commitments accordingly.

This extension will further give the opportunity for a foundation of artificial legal institutions and for their connections with informal ones. In fact the "basic norm" (Kelsen, 1967), *i.e.* the basic informal institution that provides the validity of legal systems, will be represented on top of the model of the other informal institutions. Representing the "basic norm" is in fact the crucial step for making it possible for a MAS to create and maintain by itself a legal system that is acknowledged as valid by the agents themselves.

The long term project is then to provide a three-layered model in which legal institutions, social institutions, and the socio-cognitive relations between the agents dynamically interact in order to enable institutional change and adaptation.

We have studied group doxastic attitudes and in particular belief and acceptance, from their philosophical characterization toward their logical formalization. In particular, we consider that group acceptance depends on an institutional context. We have also highlighted how to anchor these institutions in agents' mental attitudes, *i.e.* how to build them on group acceptances.

As remarked in the Section 3.3.1, group belief is deeply related to dialog. Indeed as Gilbert's group belief is the result of a consensus of group members, members need to communicate to reach this consensus state. Thus in the production of the group belief there is a strong link between group belief and communicative actions. Moreover the group belief viewed as the product of the dialog can thus characterize in some way this dialog. The aim of following chapters is thus to analysis links between group belief (and acceptance) and dialog.

Chapter 5

Theories of agent communication

5.1 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 (Searle, 1969) presented in Section 5.2, 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, mentalist accounts (presented in Section 5.3) with KQML and FIPA-ACL in particular are the most widely used to formalize agent interaction protocols. FIPA-ACL is semantically rich, and the involved concepts are quite intuitive.

Nevertheless, mentalist ACLs have 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, according to FIPA-ACL (FIPA, 2002a) 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, for example that j believes that i to be sincere and competent; 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. Therefore mentalist semantics have been criticized as being non-verifiable.. This is especially felt as being too strong in open environments with black- or gray-box agents to which we do not even want to ascribe mental attitudes. In contrast, semantics based on the concept of social commitment (Singh, 1998) (presented in Section 5.4) are 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 second chapter is to resolve the problems of mentalist semantics without loosing its benefits. In Section 5.5) we propose a novel semantics avoiding the strong hypotheses of the original semantics by "shifting" the BDI-based 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. We will show in chapters 6 and 7 how the group belief logic can be used to bridge the gap between these two approaches.

5.2 Speech Act Theory

5.2.1 Introduction

Speech Act Theory has been introduced by Austin in the 50ies in opposition to the traditional theory of language established by Frege and Russell (Frege, 1971; Russell, 1989). The latter authors were mainly interested in utterances describing the actual world. They also aim to logically describe the language and thus to ascribe a truth value to such utterances. For example the descriptive utterance "The sun is shinning" only describes the world and can be true or false only depending the actual state of the world. Other uses of language were often ignored. For example, open-questions was not viewed as a well-formed utterance.

Austin's starting point (Austin, 1962) was the study of utterances in affirmative form, singular first-person, present indicative tense and active voice, that do not describe actual world, but rather perform some action. For example he considers sentence as "I pronounce you man and wife" or "I promise that I will go to the cinema with you tomorrow". By uttering the former sentence the mayor actually performs the action of marrying the couple, whereas with the latter a father incurs a promise (and a commitment) to his son. Thus to tell is to act. In opposition to descriptive ones (named *constatatives* by Austin), such *performative* utterances do not have truth values: indeed they seem to be indifferent to the truth of their content. But as other actions, their attempt can succeed or fail for Austin performatives can be either happy or infelicitous (or unhappy). This kind of utterances are thus evaluated with *felicity conditions*.

But the dichotomy between constative and performative cannot hold because of, among other reasons, implicit performative acts such as: "I will go to the cinema with you". This utterance can be literally understood as a constative about a future action that I will perform. But it can also be interpreted as a performative equivalent to: "I promise to go to the cinema with you". Thus Austin abandoned thus this dichotomy (Austin, 1962). He extends performatives to a larger set of utterances, including constatives and performatives. The performativity becomes more generally the feature that have some utterances under some conditions to be acts. For example by uttering "the cat is sleeping",

I describe a state of the world but I also perform the action of asserting that the cat is sleeping. Austin argues that in the utterance of such sentences, three acts can be distinguished:

- the *locutionary* act, *i.e.* the action of uttering a sentence;
- the *illocutionary* act, *i.e.* the action of uttering the content with a particular illocutionary force;
- the *perlocutionary* act, *i.e.* the action performed by uttering.

For example, by uttering "the sun is shining", I perform the locutionary act of uttering the sentence "the sun is shining" (say p), but also the illocutionary act of asserting that p. Finally I also intend to cheer the hearer up, which is the perlocutionary act[1].

As expressed above an utterance can be performed with various aims, for instance to assert something or to threat the hearer. When a father tells his son: "I won't give you pocket money anymore.", he can threat him to suppress his spending money or also simply utter the fact that he will not give him money anymore. Thus the utterance does not hold alone the whole sense of what has been uttered. In particular, this utterance can have various *forces*: in this case assertive or commissive (here a threat). Contrarily to Frege, in Speech Act Theory the basic meaning unit is not the sentence but the illocutionary act, that will be represented in the sequel by $F(p)$, with F the illocutionary force and p the propositional content.

5.2.2 Five basic illocutionary forces

Searle (Searle, 1969) distinguishes five basic kinds of illocutionary forces. This distinction is mainly based on various kinds of direction of fit.

Assertive. The illocutionary point of assertive speech acts is to describe the actual state of the world. The direction of fit is therefore world to word: the words are chosen to stick to the world. The propositional content is thus a state of affairs. As examples we can cite some verbs that can have assertive performative use: assert, inform, claim, affirm, report, postulate ...

Commissive. The illocutionary point of commissive speech acts is to put by words the speaker himself in a situation in which he has to perform an action (and thus to change the world accordingly to words). The direction of fit is therefore word to world. The propositional content describes an action that the speaker will have to perform. As examples we can cite the following verbs: promise, pledge, undertake, accept, reject, agree, refuse...

[1] Searle (Searle, 1969) criticizes the last kind of act because by uttering p, an agent cannot enforce to cheer the hearer up: he can also involuntarily get him down, thus the speaker can never control the perlocutionary act, indeed it appears rather to be hearer dependent. In this sense, it is not an action.

Directive. Directives are similar to commissives in the direction of fit: the aim is that the world is altered. But in directives the requested action will be performed by the hearer. As examples of directive verbs we can cite: request, order, command, appeal, ask, supplicate...

Declarative. They correspond more or less to performatives as initially studied by Austin: by performing a declarative speech act, the speaker changes the world immediately, and the words describe also this new state of the world. This corresponds thus to a double direction of fit (world to word and word to world). As examples we can cite: declare, renounce, resign, bless, name, authorize ...

Expressive. The point is no more related to the world: it allows the speaker to express his private mental attitudes (as goals, desires, emotions...). The direction of fit is thus the empty direction: words do not describe the world, nor do they aim at modifying the world. As example we have: congratulate, blame, cheer, apologize, greet ...

5.2.3 Characteristics of illocutionary force

Searle and Vanderveken (Searle and Vanderveken, 1985; Vanderveken, 1990) have developed a taxonomy of various illocutionary forces in function of six main features.

Illocutionary point. There always exists a particular relation between the language, the propositional content and the world of the utterance. The illocutionary point is the main feature of the illocutionary force because it determines the *direction of fit* of the utterance with the world. For example in an assertion, the language has to stick to the world (there is a world-to-word direction of fit). In contrary by ordering, the speaker wants to change the world by the action of the hearer (this is a word-to-world direction of fit). The distinction between the four distinct directions of fit will be developed in the following section (Section 5.2.2).

Mode of achievement. It determines the particular way in which this force must be achieved on the propositional content. For example, with a request or an appeal the speaker gives to the hearer the opportunity to refuse to perform the requested action whereas with an order or a command, the speaker does not leave an refusal option thanks to his social and hierarchical power over the hearer.

Propositional content conditions. Some acts can be performed only on a particular kind of propositional content. For example, a promise has to be about a future course of actions of the speaker whereas the propositional content of a request must represent a future course of action of the hearer.

Sincerity conditions. By performing a speech act, the speaker often expresses one of his private mental attitudes. For example, by asserting he expresses his belief in the propositional content, by promising his intention to perform the promised action or by congratulating his happiness. To perform the speech act successfully the expressed mental attitude should hold. For instance to assert successfully the speaker has to believe the propositional content of the asserted sentence.

Preparatory conditions. Some additional conditions should also hold and be presupposed by the speaker to a performance of the illocutionary act, otherwise the act will be said defective. For example, to congratulate is to express his happiness for something good that has happened. This good fortune is presupposed for the successful performance of the congratulation, it composes thus the preparatory condition.

Degree of strength. The speaker can perform the same kind of speech act with a varying intensity degree. For example order, request and appeal are three performative verbs which aim to the performance of some action by the hearer. But order is stronger than request, itself stronger than appeal.

5.2.4 Conclusion

This operative view of the language where communication is viewed as the performance of particular actions and the link given between speech acts and agents' mental states have induced a very productive convergence of research areas such as artificial intelligence, philosophies of mind and action and linguistics. Thanks to this convergence Speech Act Theory has allowed the development of formal Agent Communication Languages.

Albeit all these advantages, Speech Act Theory has often been criticized. The most relevant criticism for our purpose is its purely monologic feature. This theory of language describes each utterance independently, regardless other preceding ones. The development of ACLs requires additional structures such as protocols or dialogue games to capture features of dialogue.

5.3 Toward an intentional approach of the dialogue: mentalist ACLs

5.3.1 Introduction

The intentional approach of dialogue is mainly based on the notion of individual intention following philosophical works of (Searle, 1983) and (Bratman, 1987). As presented in the previous chapter these works have been afterwards used in AI to propose logical frameworks to represent links between intention and other mental attitudes (Cohen and Levesque, 1990a; Rao and Georgeff, 1991; Sadek, 1992; Herzig and Longin, 2004). Similar logical frameworks have been

used to define the semantics of some ACLs. These kinds of ACLs are named *mentalist* because their semantics is defined in terms of agents mental attitudes. Among various ACLs that were developed, we only present here the two main ones: the first historically one KQML and the standard FIPA-ACL.

5.3.2 Philosophical, logical and theoretical foundations

For Grice (Grice, 1975), an utterance can have two kinds of meaning. The *natural meaning* refers to a kind of immediate relation between cause and effect such as in the sentence "These dark clouds mean rain". In contrast, Grice highlights the link between the *non-natural meaning* of an utterance and its speaker's intentions. Consider the following example:

> "When a diplomat says yes, he means "perhaps";
> When he says perhaps, he means "no";
> When he says no, he is not a diplomat.
> When a lady says no, she means "perhaps";
> When she says perhaps, she means "yes";
> When she says yes, she is not a lady.
> Voltaire (Quoted, in Spanish, in Escandell 1993.)" (Korta and Perry, 2006)

In this lighthearted example, there is a gap between the meaning of the word that the diplomat utters ("yes") and the meaning he intends to communicate, that is "perhaps". This communicative intention is oriented toward the hearer and is satisfied when it has been recognized. In order to explain how the non-natural meaning can be understood or the communicative intention can be recognized, Grice introduced the cooperation principle (developed under the form of the well-known maxims of conversation[2]):

> "Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged." (Grice, 1975)

Following this idea, the intentional approach is mainly based on the notion of individual intention: "the dialogue structure is only an epiphenomenon resulting from conversing agents' intentions (with possibly cooperation)" (Pasquier, 2005). (Grosz and Kraus, 1996) extend this idea and argue that it is possible to distinguish for every utterance of the dialogue a particular purpose that is a sub-goal of the dialogue goal.

[2]The four maxims can be summarized by:

- **Maxim of Quality**: Do not say what you believe to be false or for which you do not have adequate evidence.
- **Maxim of Quantity**: Contribution have to be as informative as it is required (and no more).
- **Maxim of Relation**: Be relevant.
- **Maxim of Manner**: Be orderly, brief and as clear as possible (without ambiguity).

	Inform(A,B,p)
Preconditions	$Know(A, p)$
Body	$MB(A, B, Want(A, Know(B, p)))$
Effects	$Know(B, Know(A, p))$ and $Know(B, p)$

Table 5.1: Action inform (Pasquier, 2005)

Following Speech Act Theory, utterances are thus considered as particular actions. Bruce (Bruce, 1975) first used AI to address the issue of the dialogue and has inspired significant contributions such as (Cohen and Perrault, 1979; Allen and Perrault, 1980). The latters develop a framework with agents endowed with the following mental attitudes: beliefs *Bel*, goals *Goal*, knowledge *Know* and mutual belief *MB*. Speech acts are represented as classical actions with preconditions, body and effects (see for example the inform act in table 5.1) and the planner used is a kind of STRIPS (Fikes and Nilsson, 1971). The principle of cooperation takes here the form of a principle of *goal adoption*: after having recognized the intention and the plan of the speaker, the hearer will adopt the goal to provide the speaker's lacking pieces of information.

The main drawback of these plan-based accounts is that they do not define a formal and well-grounded semantics. Logical frameworks have thus been developed (Cohen and Levesque, 1990c; Sadek, 2000) based on epistemic (and doxastic) and dynamic modal logics, called *rational interaction theories*. Intention is thus introduced as a primitive operator (Sadek, 1991) or as an abbreviation defined in terms of beliefs, goals and time (Cohen and Levesque, 1990a). These logics are used to define a formal semantics for ACLs, as shown in the sequel.

5.3.3 KQML

5.3.3.1 Presentation

KQML (Knowledge Query and Manipulation Language) developed as a part of the ARPA Knowledge Sharing Effort (Neches et al., 1991) was the first ACL. It was initially developed in the area of large knowledge bases and knowledge base systems distributed on internet (Finin et al., 1994). Agent technology was chosen to manage such systems and to address issues such as the heterogeneity of the knowledge bases (data structures, platforms, implementation technology). KQML is the language developed for communication between these software agents. Initially no semantics was given to various speech acts performable in KQML, which induced the emergence of various idioms without mutual understanding between these languages. We present in the following section one of these semantics.

Basically an agent (`:sender`) using a KQML language sends a message to another agent (the receiver: `:receiver`) with a particular content (`:content`) written in a given language (`:language`), and a particular performative label. The message is also understood in a given ontology (`:ontology`). Note that

Category	Names
Basic Query	evaluate, ask-if, ask-about, ask-one, ask-all
Multi-response query	stream-about, stream-all, eos
Response	reply, sorry
Generic informational	tell, achieve, cancel, untell, unachieve
Generator	standby, ready, next, rest, discard, generator
Capability-definition	advertise, subscribe, monitor, import, export
Networking	register, unregister, forward, broadcast, route

Table 5.2: Seven categories of KQML performatives (Finin, Labrou, and Mayfield, 1996)

the notion of performative does not match exactly the concept of performative verbs presented above in the Speech Act Theory: "The primary function of the performatives is to identify the protocol to be used to deliver the message and to supply a speech act which the sender attaches to the content." (Finin, Labrou, and Mayfield, 1996). A large number of performatives is defined in KQML as presented in Table 5.2, including performatives for dialogue (queries and generic informationals as tell) and performatives regulating the dialogue (responses and generators) among others (for a more detailed account of these performatives see (Labrou, 1996)).

For example, an agent `joe` could query about the price of a share of IBM stock (represented by a formula in the `LPROLOG` language) to another agent identified with the string `stock-server`, in an ontology identified by `NYSE-TICKS` by sending the following message.

```
(ask-one
    :sender joe
    :content (PRICE IBM ?price)
    :receiver stock-server
    :reply-with ibm-stock
    :language LPROLOG
    :ontology NYSE-TICKS)
```

Agent `joe` can answer afterwards using a `tell` performative as follows (Finin, Labrou, and Peng, 1999):

```
(tell
    :sender stock-server
    :content (PRICE IBM 14)
    :receiver joe
    :in-reply-to ibm-stock
    :language LPROLOG
    :ontology NYSE-TICKS)
```

5.3.3.2 Semantics: example of (Labrou and Finin, 1997)

Several semantics were developed ((Cohen and Levesque, 1990b) for example) to formalize KQML speech acts, which produced some dialects that were independent and incompatible. We present in this section one of them, that is mainly based on BDI-like representational structures, and is due to Labrou and Finin.

A complete account of this semantics is presented in Yannis Labrou's dissertation (Labrou, 1996). Performatives are described in terms of preconditions and postconditions. For a performative `performative`(i,j,φ)[3], authors denote by **Pre**(i) i's state necessary to send this performative and by **Pre**(j) j's state necessary to successfully receive and process the message (otherwise an *error* or *sorry* message will be responded). **Post**(i) and **Post**(j) hold automatically after the successful processing of the performative (*i.e.* in particular if no *sorry* or *error* has been sent in response). To describe agents' states, Labrou *et al.* uses primitive notions of Belief (BEL), Knowledge ($KNOW$), Desire ($WANT$) and Intention (INT) with a common sense meaning. For example, $BEL(i,\varphi)$ means that φ holds for agent i, and $WANT(i,\psi)$ means that agent i desires that the cognitive state described by ψ occurs in the future[4].

For example, authors give to the basic assertive performative `tell` (i expresses to j that he believes φ to be true) the following semantics:

`tell`(i,j,φ)

- **Pre**(i) : $BEL(i,\varphi) \wedge KNOW(i, WANT(B, KNOW(j,\psi)))$
- **Pre**(j) : $INT(j, KNOW(j,\psi))$ with ψ that is $BEL(j,\varphi)$ or $\neg BEL(j,\varphi)$
- **Post**(i): $KNOW(i, KNOW(j, BEL(i,\varphi)))$
- **Post**(j): $KNOW(j, BEL(i,\varphi))$

In order to tell to agent j that φ holds, the speaker i has to actually believe that φ holds and to know that j wants to know if φ holds. After this utterance, j knows that i believes φ and i knows the former.

The multiplication of various semantics has highlighted the need of a standardization in the world of ACLs. The aim of the consortium FIPA was thus to provide an ACL with one unified semantics.

5.3.4 FIPA-ACL

5.3.4.1 Presentation

FIPA-ACL is the first and only attempt to standardize an Agent Communication Language, in order to avoid emergence of plenty of semantics and independent language. The FIPA-ACL is indeed composed of three components: the

[3] This notation corresponds to a message sent by agent i (`:sender`) toward agent j (`:receiver`) with the content φ, (`:content`).

[4] *WANT* and *INT* are also about a cognitive state contrarily to *BEL* that is about an expression.

communicative act library (FIPA, 2002a), the content language (FIPA, 2002c) and some interaction protocols such as the well-known Contract Net Protocol (FIPA, 2002b). FIPA-ACL is mainly based on speech act theory: agents share messages *i.e.* they perform communicative acts (close to KQML performatives send). Their semantics is based on David Sadek's works (Sadek, 1991; Sadek, 1992). Agents have to follow some interaction protocols in order to produce a coherent dialogue. The Contract Net Protocol will be presented and discussed in the following section.

5.3.4.2 Semantics

In the standard semantics of FIPA (FIPA, 2002a), semantics is given by providing the *feasibility preconditions* (FPs) and the *rational effects* (REs) of single Communicative Acts. The former express the logical conditions to be fulfilled in order to execute the respective act, and the latter express the conditions that 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.

FIPA-ACL describes 22 Communicative Acts, that are based on four primitive ones: Inform, Confirm, Disconfirm and Request. We present here only the basic assertive Communicative Acts (Inform) and the basic directive one (Request):

$$\langle i, j, \mathsf{Inform}, \varphi \rangle$$
$$\text{FP: } Bel_i\,\varphi \wedge \neg Bel_i\,(Bel_j\,\varphi \vee Bel_j\,\neg\varphi \vee U_j\varphi \vee U_j\neg\varphi)$$
$$\text{RE: } Bel_j\,\varphi$$

The formula $Bel_i\,\varphi$ means that agent i believes that φ holds. $U_j\varphi$ denotes that agent j is uncertain about φ, but thinks that φ is more likely than $\neg\varphi$.

Directive communicative acts are defined by:

$$\langle i, j, \mathsf{Request}, \alpha \rangle$$
$$\text{FP: } FP(\alpha)[i\backslash j] \wedge Bel_i\ Agent(j, \alpha) \wedge \neg Bel_i\ Intend_j\ Done(\alpha)$$
$$\text{RE: } Done(\alpha)$$

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 $Intend_j\ Done(\alpha)$ denotes that $Done(\alpha)$ (i.e., action α has just been performed successfully) is an intention of agent j. The RE just specifies that the intended perlocutionary effect of this communication act is to get α done.

5.3.4.3 FIPA hypotheses

FIPA semantics is deeply linked to strong hypotheses on the agents. First, as presented in the Inform semantics, agents have to believe what they assert. This is the sincerity condition. Moreover additional cooperation hypotheses are added. For example the following implication is presented as a property of Feasible Preconditions persistence:

$$\models Bel_i\,(Done_\alpha \rightarrow FP(\alpha)) \qquad \text{(\textbf{Property 5} from (FIPA, 2002a))}$$

Indeed this property can also be interpreted as a cooperation hypothesis. Every agent believes that if a speaker has performed some speech act then the preconditions still holds. Under this hypothesis an agent cannot believe for example that an insincere assertion has been performed. He has to believe that other agents are cooperative. Moreover this property can also be applied to the speaker himself. In this case (and due to introspection properties) Feasible Preconditions actually hold. It follows that an agent cannot perform an insincere assertion, he cannot lie: he has to be both sincere and cooperative.

5.3.5 Advantages and limitations of mentalists approaches

FIPA-ACL is semantically rich, and the concepts involved are quite intuitive. Its standardized semantics is indeed close to speech act theory and is thus philosophical well grounded. Using mental attitudes in the definition of the semantics gives also a strong predictive power to this account. In particular it allows to infer the communicative intention of the speaker. Nevertheless, FIPA-ACL has a feature that has often been criticized in the literature (Singh, 2000; Fornara and Colombetti, 2002), viz. that the semantics of communication acts 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. Moreover this constraint imposes a kind of internal architecture for agents: they have to be able to manage mental attitudes and to do inferences. This is a strong limitations against mentalist ACLs because it imposes to designer a to use a particular kind of cognitive agents.

Following these criticisms a new trend of ACLs has been developed following Singh's work (Singh, 1998).

5.4 Social approaches: commitment-based approaches

5.4.1 "ACLs: rethinking the principles" (Singh, 1998)

Singh (Singh, 1998) notices that Agent Communication Languages have been developed only for proprietary multi agents systems. Indeed no effort has been made to allow mutual understanding. But the development of internet would favor the interaction between heterogenous agents from various MASs. Moreover languages presented above are mainly based on the assumption that agents were designed as cognitive agents with mental attitudes and inference capacities. Once again this hypothesis is too strong for open multi-agents systems with heterogenous kinds of agents. Singh argues thus that principles of ACLs should be rethought and proposes a new stream where languages are no more based on the intentional aspect of communication but rather on its public and social aspect.

Based on that he considers how ACLs should address the issue of the meaning of the communication. Many elements should be taken into account to manage this question. Afterwards he considers designers point of view in front of these ACLs.

Firstly, every communicative act can be interpreted from three *perspectives*: the speaker's (or message sender's), the hearer's (or receiver's) and the public's. Mentalist ACLs are focused on the first two. But Singh argues that mental attitudes should only be kept as a tool for some designers to specify agents behavior. Agents should be considered only as black boxes able to perform communicative acts in the ACL's definition. Thus communication meaning should only be interpreted from a public and social perspective due to agents' heterogeneity.

In addition, the *type of meaning* should also be taken into account: it can be personal or conventional. With personal meaning, an communicative act can be interpreted differently according to the agent. Typically the assertion of the intention that the hearer performs some action can be interpreted by some agents as a directive speech act, whereas it could be interpreted simply as an assertion by another one. Singh advocates a conventional type of meaning. As language is typically a convention system, meaning of every communicative act should be conventional and thus public and common to all agents.

The fact that the assertion of an intention can be interpreted as a directive induces an impoverishment of illocutionary forces taken into account by mentalist ACLs: for example in KQML, every act is a kind of assertive (`tell` in KQML). Singh argues that communication between autonomous agents needs seven kinds of communicative actions: assertives, directives, commissives, permissives, prohibitives, declaratives and expressives. ACLs as FIPA-ACL or KQML only manage in general assertives and directives. It could be possible to add other categories but only thanks to lot of efforts. Singh advocates that ACLs should be extensible enough to integrate easily new communicative

primitives.

Finally as presented above, mentalist ACLs need strong hypotheses on agents, such as sincerity and cooperation. Otherwise these languages cannot be applied in practice. Singh remarks that in open systems these hypotheses cannot be imposed because agents are heterogenous. Thus no particular kind of internal architecture can be ensured, and dialogues can be based on opposition or conflict between agents in particular in the case of negotiation, persuasion or business dialogues. ACLs should thus not be based on above constraints.

From the designer's standpoint, ACLs should not be an additional constraint on the design of the agent. Designers should be free to manage agents' behavior with their preferred tools and should not be forced to deal with mental attitudes as imposed by mentalist ACLs.

To summarize, for Singh (Singh, 1998), a new trend of ACLs is needed with new requirements. Meaning of utterances should be socially and conventionally grounded. The language should be simple and expressive enough but also easily extensible. Other authors such as (Fornara and Colombetti, 2004) require the verifiability of the dialogue adequation to the protocol managing the sequence of speech acts. This verifiability is mainly due to the public character of speech acts. To follow these requirements, Singh advocates the use of the notion of social commitment to define speech act semantics.

5.4.2 Discussion around the notion of commitment

Commitment has various senses in the AI literature. First of all, it is important to make a distinction between internal commitment and social commitment. Internal commitment "refers to a relation between an agent and an action" (Castelfranchi, 1995). In particular, commitment is used in this sense in the famous slogan "Intention is choice with commitment" (Cohen and Levesque, 1990a). It captures the persistence of an agent's choice to perform an action. The agent can be either an individual agent or a group of agents (handled as a whole). This entails a subdistinction between personal and collective commitment. The collective commitment captures the persistence of a group of agents' collective choices. It should not be confused with social commitment which refers to a particular relation between agents. In this section we focus on the study of social commitment. The personal commitment is taken into account in our framework *via* the notion of intention defined in chapter 3.

In Castelfranchi's view, social commitment is a relation between two agents only about an action. (Walton and Krabbe, 1995) have introduced another kind of social commitment: propositional social commitment (where propositions and actions are different entities). We describe both notions in the sequel.

5.4.2.1 Commitment in action

Characterization. Social commitment is a relational notion. That means it links at least two agents: the agent who is committed (the *debtor*) and the agent to whom the debtor is committed (the *creditor*) (Fornara and Colombetti,

2002). A third part can be involved in a commitment: the *witness*. Castelfranchi proposes the following definition of social commitment of the debtor i to the creditor j w.r.t. the action a:

> "[i] and [j] mutually know that [i] intends to do a and this is [j]'s goal, and that as for a [j] has specific rights on [i] ([j] is entitled by [i] to a)." (Castelfranchi, 1995, p. 3)

In opposition to Singh's account (Singh, 1991) and works about ACLs (Colombetti, 2000), Castelfranchi considers that social commitment (in action) is not primitive but can be defined in terms of agents' mental attitudes with additional deontic concepts. In the sequel we will follow Castelfranchi on his reductionist view of commitment.

This characterization will be discussed in Section 7.3. We can nevertheless make some preliminary remarks. First, the mutual knowledge on i's intention is needed, which logically implies i's intention. But as Castelfranchi admits himself, the actual intention of i to perform a is neither necessary nor sufficient for his social commitment to do a: the entailment link between social commitment and individual intention only holds if the agent is honest. We could thus weaken the definition by substituting mutual knowledge with a notion capturing only the public feature of i's intention: if i has a social commitment toward j to do a then it must be public for both agents that i has the intention to do a. This notion of public ground should be designed in a way such that (1) the public ground of an attitude does not imply automatically that this attitude holds actually and that (2), in the case where agents are honest, it should entail mutual knowledge.

Second, the action a to which i is committed should be a goal of the creditor j. We argue that this hypothesis is also too strong in general. We can compare this to the Speech Act Theory (Vanderveken, 1990, p. 182–183) by considering commissive acts. On the one hand, Promise produces a commitment from the speaker to the hearer and needs that what is promised is a goal of the interlocutor, but on the other hand Threat has the same social result (creation of a commitment) but needs in contrary that the object of the threat is not a goal of the agent (or is contrary to one of his goals). Thus, we could distinguish between what we call desirable social commitment (when the fact that a be performed is a goal of the creditor), and undesirable social commitment (when the fact that a be performed is not a goal of the creditor). Social commitment *à la* Castelfranchi corresponds to desirable social commitment. We do not need to define social commitment in such a restrictive way as we will show in the sequel.

Finally Castelfranchi introduces deontic aspects of social commitment with a vague notion of right. This point would benefit from a formal characterization of commitment. Note that as highlighted by (Carey, 1975) commitments do not imply an obligation *stricto sensu*. In particular Royakkers (Royakkers and Dignum, 2000) introduced rather a notion of directed obligation. When an agent incurs a commitment toward another agent, he does not have a general obligation but rather a kind of obligation directed toward the other agent.

Satisfaction of social commitments. Once a commitment has been incurred, it is important to characterize when it is fulfilled. (Castelfranchi, 1995) distinguishes two kinds of satisfaction: subjective satisfaction (when j believes that the action has been performed) and objective satisfaction (when the action has actually been performed). Objective satisfaction is clearly necessary. If j believes that a has been done whereas a has not been performed yet, i stays with risks of blame hanging over him (by ongoing rights of j). But it is not sufficient. j needs to be aware of the objective satisfaction, or more precisely, if i can inform j of the commitment satisfaction, he should do it.

Fornara and Colombetti (Fornara and Colombetti, 2002) among other authors (Flores, Pasquier, and Chaib-draa, 2007; Mallya and Singh, 2007) improve this account by adding additional states to the commitment life-cycle of ACL semantics: each speech act is characterized by the commitment in a particular state it produces.

5.4.2.2 Propositional commitments

Walton and Krabbe (Walton and Krabbe, 1995) follow Hamblin's work (Hamblin, 1970). The latter addressed the issue of the argument evaluation in the context of formal dialogues and used already commitment as the key notion, but without studying it in depth. Walton and Krabbe propose thus a thorough analysis of commitments, identify both kinds of commitments (in action and propositional) and use the latter to formalize dialogues.

The general form of a commitment in action is: "[i] is committed to [a]-ing" where i is the subject of the commitment and a its object. In opposition to Castelfranchi's analysis of the concept of commitment itself in relation to mental attitudes, Walton & Krabbe consider commitments in action in relation with associated *imperatives*: to know the object of the commitment is to know what should and should not be done to stay close to this commitment. They classify also commitments in function of the sanction associated to them.

Consider the two following examples to grasp the distinction between propositional commitment and commitment in action:

EXAMPLE. *On the evening, John asks Mary: "Where will you put the garbage for me to take out tomorrow morning?" Mary answers: "Behind the door, as usual." (Walton and Krabbe, 1995, p. 22)*

This answer is indeed a promise from Mary to John to put the garbage behind the door. She has thus incurred a commitment in action to put the garbage there.

EXAMPLE. *John is about to take the garbage out. He asks Mary: "Where did you put the garbage for me to take out tomorrow morning?" Mary answers: "Behind the door, as usual." (Walton and Krabbe, 1995, p. 22)*

In this second case, Mary has not promised to perform any action herself. But she has nevertheless incurred a propositional commitment on the state of affairs that the garbage is behind the door. She cannot, once this commitment

incurred, deny the fact that she has performed this assertion or assert that the garbage is elsewhere without running the gauntlet. Moreover she has to defend her commitment when it is attacked by other agents' arguments (for example that she has not performed such assertion). Thus she is also committed in some way to other actions (holding that P, defending if attacked ...) that are only dialogical. Walton and Krabbe give thus the following definition of a propositional commitment:

DEFINITION.*[(Walton and Krabbe, 1995, p. 23)]* Propositional commitment *is*

(1) a kind of action commitment whose

(2) partial strategies assign dialogical actions that

(3) center on one proposition (or a formulation thereof).

Moreover Walton and Krabbe point out that in the first example Mary has also incurred some propositional commitments in addition to her commitment in action to put the garbage behind the door. In particular, she is committed not to deny her promise and to defend the fact that she has promised. The link between these two kinds of commitment will be analysed more deeply in the sequel.

5.4.3 Description of commitment-based ACLs

Following Singh's manifest introduced in (Singh, 1998), some accounts aiming at given a formal social semantics of ACLs have emerged in the last decade. We present in this section the main contributions of Colombetti *et al.* and Chaibdraa *et al.*. Contrarily to Castelfranchi (Castelfranchi, 1995), they consider commitment as a primitive. After having characterized it in terms of its life-cycle automata, they define speech act semantics thanks to various states of commitment.

5.4.3.1 Colombetti *et al.*

In the sequel among the large literature produced by Colombetti *et al.*, we consider as reference one of their latest papers (Fornara, Viganò, and Colombetti, 2007). We will also refer quickly to some extensions such as conditional commitments or commitment with deadlines.

Commitments. A commitment is basically characterized by a *state*, a *debtor*, a *creditor*, and a *content* and denoted by:

$$C(state, debtor, creditor, content) \tag{5.1}$$

This notation means that the debtor has a commitment toward the creditor at a particular state on some content. The content is represented by a *temporal proposition* (Fornara, Viganò, and Colombetti, 2004). A temporal proposition

(TP) is a statement (that means a proposition, *i.e.* a state of affairs, or an action to perform) related to a time interval with two modes of validity ($\exists$ and $\forall$). It can take three truth values : undefined, false or true. For example, the proposition TP(φ, [t_1,t_2],$\exists$,state) stays undefined as long as φ is false and the current instant is in the interval. It becomes true as soon as φ is true and false if the current instant is later then t_2 without φ being true in [t_1,t_2]. This formalization in terms of a temporal proposition allows in particular to manage deadlines in commitments.

A commitment can be in various states as described by the commitment life-cycle described in Figure 5.1 quoted from (Fornara, Viganò, and Colombetti, 2007)[5]. State transitions are triggered by the performance of specific institutional actions, *i.e.* actions able to manage institutional facts (such as commitments) or by the changes of the content truth value. When an agent i demands or requests another agent j to perform some action, both agents are in a particular social relation, in which j is requested to commit itself to action. Colombetti represents it by an **unset** commitment from j to i about the action[6]. This means that c created himself a commitment of which he is the creditor. This commitment is created by the function *makeCommitment*. It can be **canceled** (*e.g.* if j refuses to commit to perform the action) with action *setCancel* or becomes **pending** (*e.g.* if j accepts the request) with action *setPending*. If the commitment content becomes true (resp. false), the commitment is **fulfilled** (resp. **violated**). An extension of this life-cycle has been proposed in (Fornara and Colombetti, 2007) to take into account an additional concept of sanction (see Section 5.4.3.2 for more details about the notion of sanction). Two states are added to the cycle representing **extinguished** commitments (that have been repaired by performing the $d{-}sanction$) and **irrecoverable** commitments (if the violation has not been remedied).[7]

Note that this automaton is independent from the type of the content, and in particular fulfilment and violation conditions are the same for a proposition and an action. We will discuss this point later. Moreover institutional actions, *makeCommitment*, *setPending*, *setCancel*..., are defined independently

[5] Note that this version of commitment does not take into account conditional commitments. For conditional commitments (Fornara and Colombetti, 2004), a state *active* is added between pending and fulfilled/violated to represent pending commitments with true condition. In this dissertation we consider only unconditional commitments (which can be viewed as conditional commitments with a tautologous true conditions; the active and pending states are thus merged.) Commitments are then represented under the form (with an additional deadline):

$$C(state, debtor, creditor, content|condition[, timeout])$$

.

[6] This unset state has also been called precommitment in (Fornara and Colombetti, 2002).

[7] Commitments take thus the form:

$$C(state, debtor, creditor, content, d{-}sanctions, e{-}sanctions, source)$$

$d{-}sanction$ is the action that the agent who has violated the commitment can perform to remedy the violation, $e{-}sanction$ is the action that the empowered agent can perform against the violation. Colombetti *et al.* argue that information about the *source* (norms ...) of the commitment is needed to handle it and its sanctions.

from speech acts: they are defined at the level of the system managing commitments. They are used at a second step to define high level speech acts semantics.

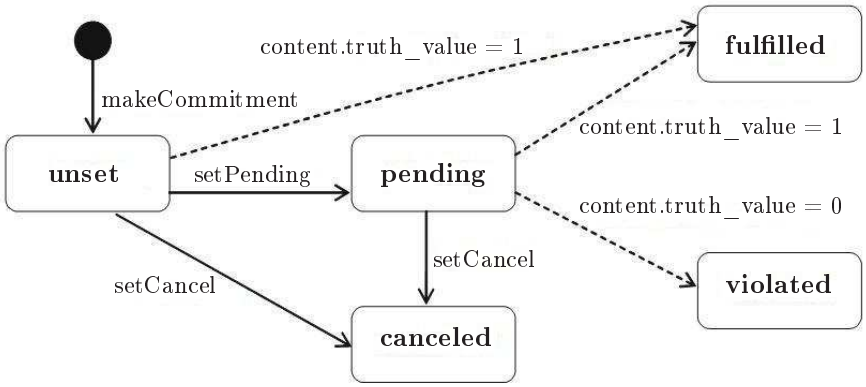

Figure 5.1: Commitments life-cycle (Fornara, Viganò, and Colombetti, 2007). The dotted lines represent events such as content truth value changes.

ACL. Following Searle (Searle, 1969), Colombetti *et al.* define speech act as institutional actions (Colombetti and Verdicchio, 2002). When an human being performs some utterance or when an artificial agent sends some message to another agent, given particular contextual conditions, this counts as the performance of a communicative act. Moreover they consider that those communicative acts can be characterized by the institutional facts that they bring about. These institutional facts depend of course on the particular institution under consideration. They adopt the working hypothesis that there always exists a primary or *core institution* that manages the basic institutional facts that are fundamental for every interaction, *viz.* social commitments.

Thus their project is to define speech act effects on the institutional layer thanks to the two institutional actions presented above: "make commitment" ($\mathsf{mc}(i, j, \varphi)$ for short) which creates a new commitment in an unset state with i as creditor, j as debtor and φ as content and "set commitment" ($\mathsf{sc}(i, state)$) which changes the state of the commitment i toward state *state*. Actions *setCancel* and *setPending* are only particular instances of the function sc. The authors give to these two basic actions the following semantics in terms of preconditions and postconditions (represented in Object Constraint Language OCL (Object Management Group, 2003)). We present here only the semantics of the action mc:

- name : $\mathsf{mc}(debtor, creditor, content)$
- pre : not Comm.allInstances → exists(c|c.debtor = debtor and c.creditor = creditor and c.content = content)

- post : Comm.allInstances → exists(c|c.state = unset and
  c.debtor = debtor and c.creditor = creditor and c.content = content)

The precondition says that an agent cannot incur a commitment if another one already exists in the set of commitments (*i.e.* another one with the same debtor, creditor and content). The postcondition says that the action creates a commitment.

Using the two basic commitment acts mc and mc, the authors choose to give a semantics only to four of the five illocutionary forces existing in speech act theory, arguing that expressive speech acts are not relevant for artificial multi-agents systems. As an illustration of Colombetti's semantics, we present in the sequel the definition of assertive and directive speech acts, under the form given in (Colombetti, Fornara, and Verdicchio, 2002)[8], that we have simplified due to the fact that we consider commitment as unconditional.

Assertive: Inform. Inform is taken as the representative of the class of assertives. From an observer's point of view, when i informs j that φ then the only thing we can say is that i has incurred a commitment on the truth of the proposition φ. The semantics of the inform speech act is thus defined by:

$$\langle i, \mathsf{Inform}, j, \varphi \rangle := \{c \leftarrow \mathsf{mc}(i, j, \varphi); \mathsf{sc}(c, \mathit{pending})\} \tag{5.2}$$

After an inform act, a pending commitment of i toward j thus holds, which can be represented in our dynamic logic by (without giving a particular semantics to the "commitment" predicate $C(\mathit{pending}, i, j, \varphi)$ in logical framework)

$$\mathit{After}_{\langle i, \mathsf{Inform}, j, \varphi \rangle}\ C(\mathit{pending}, i, j, \varphi) \tag{5.3}$$

Directive: Request. Request is taken as the representative of the class of directives. In the general case, by requesting an agent j to perform some action α, it would be too strong to say that i has directly created a pending commitment of j with i as creditor. Even when there exists strong hierarchical relations between agents, as in the army, a creditor cannot create a commitment by performing a strong directive as an order, but only that he has created an obligation that the debtor should respect. To represent the particular relation existing between both agents after the performance of a directive, Colombetti *et al.* use the *unset* commitment state. Thus the performance of such an illocutionary act counts as the creation of an unset commitment:

$$\langle i, \mathsf{Request}, j, j{:}\alpha \rangle := \{\mathsf{mc}(j, i, j{:}\alpha)\} \tag{5.4}$$

After a request act, an unset commitment of d toward c thus holds, which can be represented in our dynamic logic by:

$$\mathit{After}_{\langle i, \mathsf{Request}, j, j:\alpha \rangle}\ C(\mathit{unset}, j, i, j{:}\alpha) \tag{5.5}$$

[8] Although the form has slightly changed along different articles the main idea remains the same.

Several aspects of the Castelfranchi *et al.* account of commitments, their life-cycle and their use to define an ACL semantics will be discussed in the sequel.

Discussion. First, Colombetti *et al.* manage propositional commitments and commitments in action by the same mechanisms. But as Walton & Krabbe show in (Walton and Krabbe, 1995), the two kinds of commitments are different in many respects. In particular the life-cycle automata presented above is dedicated to both kinds of commitments, but it does not seem to be fully adapted to propositional commitments. For example the state *unset* is not reachable by propositional commitments because the only speech act creating a propositional commitment is the Inform speech act and it creates automatically a *pending* commitment. Moreover the relevance of the *unset* state is questionable for propositional commitments. An unset commitment on a proposition would be the result of a request that the debtor become committed on a state of affairs. But as a commitment can only be incurred by performing an action, this request must be a request to perform the action to be propositionally committed. For example when a policeman requests a suspect person to avow that he has killed his wife, it is more intuitive to represent this state of affairs by an unset commitment to perform the action of avowing rather than by an unset commitment to be committed to the fact that he has killed his wife. Moreover following Speech Act Theory, the content of speech acts with a directive (as request) or commissive illocutionary force can only be an action.

Secondly, for Colombetti *et al.*, a commitment (on proposition or on action) is fulfilled (resp. violated) as soon as its content becomes true (resp. false). But following Walton and Krabbe (Walton and Krabbe, 1995), propositional commitments represent what has been asserted in a dialogue, and have to be defended by arguments when challenged[9]. Thus such a commitment is not related to the truth of its content. Its state is rather determined by the agent's capacity to maintain it. A speaker can utter a false proposition while being convinced it is true. We cannot say that he has violated his commitment if in the sequel of the dialogue he stays coherent with it. We cannot say either that a commitment is fulfilled if its content is true but its debtor retracts what he has said and concedes the contrary.

For example, several centuries ago, most scientists believed that earth was flat. Let us consider a debate about earth boundaries and what there is on the edge. By asserting that the earth is flat as a premise of an argument X, agent a incurs a commitment on the fact that the earth is flat. Following Colombetti's characterization, we can say that his commitment is violated. But in fact, neither a nor his interlocutors can detect that this commitment is violated and moreover they will believe that it is fulfilled.

Consider a second example. John tells Mary: "I believe that it will be sunny

[9]Remark that we use here vocabulary used in (Gaudou, Herzig, and Longin, 2006b) in the case of the formalization of Walton & Krabbe's persuasion dialogue. But we are not limited here to that particular kind of dialogue type.

tomorrow". John has incurred a propositional commitment to the proposition: "I believe that it will be sunny tomorrow". One more time, this commitment cannot be verified by agents other than John: Mary cannot know John's private mental attitudes (we recall that this is one of the most important criticisms against mentalist ACLs).

According to Colombetti, the fulfilment condition of those two commitments can be verified only from an objective point of view, *i.e.* the one of an omniscient agent. We consider that this is problematic in multi-agent systems where each agent must be autonomous: for the interaction with other agents should not depend on a central agent. Thus rather than an objective concept we will see commitment as a public concept and keep the coherence condition rather than the truth value condition to determine the state of the commitment. We consider here that a public concept is a subjective concept in the sense that each agent can determine alone its truth value. Contrarily, an objective concept is objective, *i.e.* its truth value is related to the actual one of an other proposition. As the agents' perception of their environment is neither sound nor complete, they cannot determine the truth value of such a proposition. Consequently, concerning fulfilment and violation condition a distinction between commitments about propositions and commitments about actions should be made.

Considering previous remarks we argue that we need two distinguished commitment life-cycles depending on the content type (this point will be developed in Section 7.3).

The last point is that Colombetti does not explore relations existing between commitments in action and propositional commitments. For example, by promising : "I promise to take out the garbage", John is incurring a commitment of taking out the garbage. Moreover, following Speech Act Theory (Vanderveken, 1990), he is also expressing at least that he has the intention to perform the action to take out the garbage. Thus he is also committed on this proposition. He can thus suffer sanction if he does not perform this action (such as reproaches by his wife), but also if he merely utters that he does not have this intention (in which case the sanction would be at the dialogue level: he would appear to be incoherent and thus untrustworthy). We can observe two kinds of violation corresponding to the two kinds of commitments. As we will show in Section 7.3.2, by incurring commitment on action[10], an agent incurs *de facto* some propositional commitments.

We have discussed above the *unset* state for propositional commitments, but we can make some remarks about action commitments. As we said above, Colombetti represents the particular social relation between two agents that results from the request by agent i toward agent j to perform an action, by a commitment of j toward i with unset state. As in (Fornara and Colombetti, 2002), we consider that "it is impossible for an agent to create a commitment of which some other agent is the debtor". We argue that after a request of i it is

[10]We consider here that agents only incur commitments intentionally by performing speech acts. We do not take into account commitments incurring automatically, *e.g.* by representatives or due to social position (see (Walton and Krabbe, 1995, p. 32), for a detailed account of the ways of incurring commitments).

doubtful that agent j is a debtor of any kind of commitment. We consider that in this case, it is only i who is committed to something, and as we will show in the sequel, he is committed to his intention that j performs the action.

5.4.3.2 Chaib-draa *et al.*

Chaib-draa *et al.* (Pasquier, Bergeron, and Chaib-draa, 2004) propose a dialogue system based on the notion of social commitments to represent the social layer managing agents interaction and dialogue games as dialogue units manipulating those commitments.

Commitments. Chaib-draa *et al.*'s approach of commitment is close to Colombetti's. Commitments are represented by a primitive operator. Contrarily to Colombetti's their content is restricted to actions, thus they do not take into account propositional commitments. Moreover in its basic version, commitments are represented with explicit notions of sanctions and deadline. The authors introduce sanctions for both agents involved in the commitment. A commitment is thus a 6-tuple: an agent i has a commitment toward agent j to perform an action α at time t. s_i and s_j describe sanctions that can be applied to both agents: if agent i cancels or violates the commitment, he will be under the sanction s_i, whereas agent j will be under sanction s_j if he tries to withdraw the commitment once he accepted it.

Commitments are thus represented by the formula (Pasquier, Flores, and Chaib-draa, 2004):

$$C(i, j, \alpha, t, s_i, s_j) \tag{5.6}$$

Sanctions. In other approaches based on commitments to represent social links and interaction between agents, commitments are often supposed to be respected. Indeed only few is said about the case of violation of these commitments (see for example (Fornara and Colombetti, 2007) in particular). But commitments are a much more flexible notion than obligation: it can be retracted or canceled. Chaib-draa *et al.* propose thus in (Pasquier, Flores, and Chaib-draa, 2004) a detailed account of sanctions allowing for the possible evolution of commitments into states such as **violated** or **canceled**.

The authors base their account on the theory of social control (Martindale, 1978): "The concept of social control originally denoted the capacity of a group or society to regulate itself and to secure coherency and unity in social life" (Pasquier, Flores, and Chaib-draa, 2004). Authors consider particularly two main elements: sanctions and punishment policies.

A sanction is characterized by three dimensions:

- its **direction**: *positive* (or reward) as incentive to follow particular behavior, and *negative* as punishment against violations);

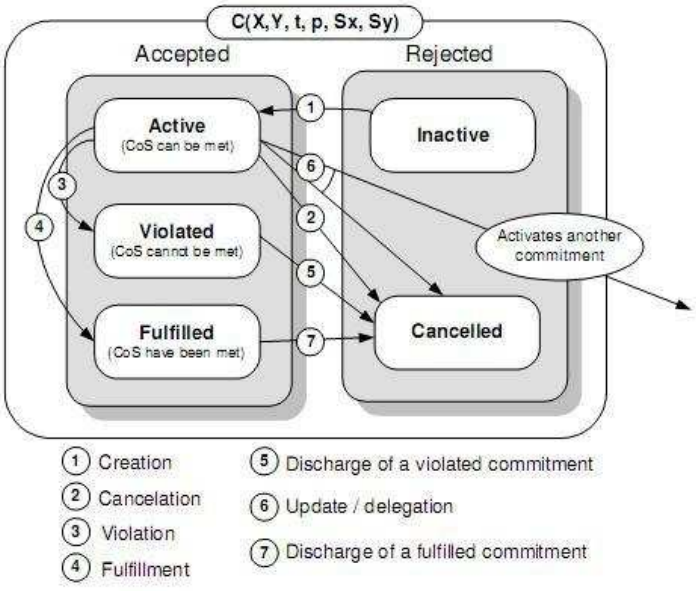

Figure 5.2: Commitments life-cycle (Pasquier, Flores, and Chaib-draa, 2004)

- its **type**: *material*, *i.e.* violence, repairing actions or financial; *social* (Posner and Rasmusen, 1999), *i.e.* sanctions affecting trust, credibility or reputation; *psychological*, *i.e.* sanctions inducing unpleasant emotions as guilt or shame;

- its **style**: *implicit*, *i.e.* private, unilaterally sanctions and *explicit* that are public.

Sanctions can also be split into two parts by distinguishing *a priori* (statical) sanctions from *a posteriori* (dynamical) sanctions.

The punishment policies (Beccaria, 1963; Bentham, 1970) determine the type and the magnitude of sanction to impose on agents, relative to designer's desiderata to fight against violations. (Vold, Bernard, and Snipes, 2002) present the five basic punishment policies: incapacitation, rehabilitation, restoration, deterrence and retribution. Chaib-draa *et al.* choose only to take into account the two latter ones, because in incapacitation, rehabilitation and restoration the choice is led only by the style and the type of the sanction, whereas deterrence and retribution focus on the strength of the sanctions. Deterrence focus is to avoid commitment violation. It aims at discouraging agents to violate their commitments by imposing strong and explicit sanctions; sanctions should be much more worth than the possible benefit of a violation. In contrary in retribution the sanction should have the same value as the violated commitment.

It is important to note that with deterrence policy less commitments will be incurred: indeed agents will be more careful and inactive, due to prohibitive punishments. In contrary with a retribution policy more commitments will be incurred (but more will be violated); agents will be freer to interact: in particular in an example of MAS for business much more transactions will be run up. Deterrence appears to be less effective (Polinsky and Shavel, 1998). But the choice of the punishment policy only depends on the designer's desiderata.

ACL. In Chaib-draa *et al.*'s account, the commitments are managed (discussed, created, canceled...) through dialogue games. Dialogue games are a particular structure of the dialogue that is more flexible than protocols. A dialogue game is defined by:

- Entry conditions (E)
- Success conditions (S)
- Failure conditions (F)
- Dialogue Rules (R)

These conditions and rules are defined in terms of commitments (Maudet and Chaib-draa, 2002). (E) represents the conditions imposed on the set of commitments to enter in this dialogue game. (S) (resp. (F)) are conditions determining the output state (Success, resp. failure) of the game. Moreover by entering in a particular dialogue game, an agent commits itself to observe the rules (R) of this dialogue. These *dialogical commitments* are specific to the game and do not hold in another one.

Several dialogue games are defined as for example the request game corresponding to the "negotiation" of a commitment to some action α. The initiator i requests j to perform action α, afterwards the latter can either accept or refuse the request. It is formalized with the following rules and conditions (note that sanctions are omitted for clarity sake):

$$
\begin{aligned}
E_{rg} :& \neg C(j,i,\alpha,t_i ni) \text{ and } \neg C(j,i,\neg\alpha,t_i ni), \forall t_i ni, t_i ni < t_f in \\
S_{rg} :& C(j,i,\alpha,t_f in) \\
F_{rg} :& \neg C(j,i,\alpha,t_f in) \\
R_{rg} :& \text{ 1) } C_g(i,j,request_{d_1}(i,j,\alpha),t_f in) \\
& \text{ 2) } C_g(i,j,request_{d_1}(i,j,\alpha) \Rightarrow \\
& \quad C_g(j,i,accept_{d_2}(j,i,\alpha) | refuse_{d_3}(j,i,\alpha),t_k),t_f in) \\
& \text{ 3) } C_g(j,i,accept_{d_2}(j,i,\alpha) \Rightarrow C(j,i,\alpha,t_f),t_f in) \\
& \text{ 4) } C_g(j,i,refuse_{d_3}(j,i,\alpha) \Rightarrow \neg C(j,i,\alpha,t_f,t_f in)
\end{aligned}
$$

These dialogue games can be combined by relations of sequence, choice or embedding as proposed in (Labrie, Chaib-draa, and Maudet, 2003; Reed, 1998; McBurney, Parsons, and Wooldridge, 2002) among others.

Their account is implemented in a software called DIAGAL (Pasquier, Bergeron, and Chaib-draa, 2004). DIAGAL claims to be a very useful and flexible dialogue manager that can moreover manage different intensity degrees of illocutionary force and the social context of interaction (*e.g.* hierarchy relation between agents that can reduce the choice of an agent).

Discussion. Chaib-draa *et al.* propose an account using commitments and dialogue games to manage the dialogue. We can remark that dialogue games represent a protocol encoded by the rules (R) with additional conditions (E), (S) and (F). The more interesting point about dialogue games is the management of shift or embedding of dialogue games. But we have to highlight that in this account, commitments are used as states of the protocols automata or as entry and output conditions. They appear as a kind of tags representing a state of the dialogue rather than the concept of commitment as described by Castelfranchi. Indeed it lacks links with other notions managing the interaction between agents.

5.4.4 Benefits and limits of social approaches

The use of social notions such as commitment to describe ACLs semantics settles the problems of the mentalist approaches. In particular, the semantics is defined at the social layer, allowing the verification of the dialogue and its compliance with interaction protocols. Moreover, this kind of semantics avoids strong hypotheses on agents. As only the public and social aspect of communication is taken into account, no sincerity or cooperation hypotheses are needed. Moreover nothing is imposed on the internal architecture of the agents: as semantics is not defined in terms of agents' mental attitudes, designers can feel free to choose their internal architecture.

As highlighted by (Dignum and van Eijk, 2007), whereas lot of works are developed based on commitments to manage agent interaction, only a few investigations about commitments themselves exist. Indeed no formal characterization has been given to the notion of commitment and its link in particular with agents' private mental attitudes. For example, when an agent is committed to perform α_1 and α_2, is he necessarily committed to perform α_1. Does a propositional commitment to $\varphi_1\varphi_2$ implies a commitment to φ_1? Moreover, as mental attitudes are not taken into account in this account, nothing is said about the internal management of the commitment (within the agent architecture).

5.5 Toward the need of an alternative: ACLs, rethinking the principles (again)

Singh (Singh, 1998) successfully initiated a reexamination of the principles of ACLs semantics. He proposed to take into account the social aspect of the conversation and to anchor speech act semantics in this social layer, in particular by introducing the notion of commitment. Many authors have followed his approach, describing speech act semantics in terms of commitments (and only in

terms of commitments). Although this approach is interesting, we argue that it remains insufficient. It is clearly well adapted to the case of commissives or to manage e-business transactions. But it does not address the management of these commitments and of the communication in general by the agent himself. To consider that the agent has to follow blindly a communication protocol decreases the attractiveness of the agent technology. Contrarily, if we consider that agents should be able to infer or deliberate to be more adaptative, we should give them the means to make these inferences. The most efficient way to handle the interaction of one agent with other ones is certainly to give him mental attitudes and inference capacities. If an agent believes that the hearer has incurred a commitment, he can infer for example that (at least publicly) he has the corresponding intention, and that he has a kind of obligation.

We will present in the sequel an alternative approach to ACLs semantics. In the tradition of Speech Act Theory and following mentalist approaches, we consider that speech acts are deeply linked to mental attitudes. But with social approaches, we agree on the social and public feature of the dialogue. We aim at bridging the gap between both approaches to ACLs. We thus argue — and this is the key point of this chapter — that ACLs semantics can be defined in terms of public mental attitudes. We show in the sequel how we can use the operator of group belief defined in the preceding chapter to take into account what is public in the dialogue and to characterize ACLs semantics.

5.5.1 Discussion about underlying hypotheses

Despite their differences all works on ACLs presented above are grounded on a common and very strong hypothesis: that is the communication between agents is perfect. That means that communication is complete (if an utterance has been performed by a speaker, the hearer is aware of this utterance, and there is no loss of information in the transmission) and sound (a hearer cannot believe that the speaker has uttered something if it is not the case). This hypothesis is indeed really strong because it removes every possibility of misperception and incomplete perception: this avoids the problem of grounding a common belief such as in the Byzantine Generals problem (Fagin et al., 1995). This hypothesis is indeed unrealistic when it is applied to dialogues between human beings, but it remains defendable in the case of artificial agents thanks to secured interaction protocols used for the hardware communication.

Traum (Traum, 1994) has addressed this issue by studying the grounding process in dialogue, *i.e.* mechanisms used to put some propositions on the common ground of a conversation.

5.5.2 Clark and Traum's dialogue model: grounding theory

In his doctoral dissertation David Traum exposed a computational model of the grounding[11] process (Clark and Schaefer, 1989). He defined it as the process describing "how [...] common ground is engaging in conversation". In particular, this problem is really hard in the case of human-human conversation in Natural Language. In this case, there is no complete mutual understanding between individuals because either the communication is not perfect (in the sense that a message can be lost or its meaning can be altered during the transmission), or the meaning of sentences is not unique or completely unambiguous. In this kind of communications, rather than reaching complete and perfect mutual understanding (*i.e.* a mutual belief in the sense of (Halpern and Vardi, 1989)) of the speaker's communicative intention and of the message only a partial comprehension is enough, where partial has to be understood in the sense of the "grounding criterion" (Clark and Schaefer, 1989).

Traum proposed a protocol for grounding what has been uttered by conversation participants. It is based on some particular speech acts, named grounding acts, such as Acknowledge, Repair... After performance of such a protocol, the utterance of the speaker becomes grounded for both participants and thus it becomes mutually believed that the speaker proposed something (*cf.* Figure 5.3). Each utterance is attempted to be grounded in this way, and is added to speaker's proposal in the Mutual Belief. In case of agreement on a proposal, it becomes shared for both participants.

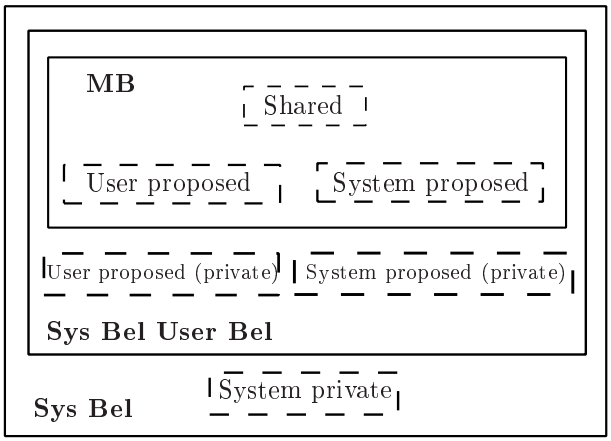

Figure 5.3: Belief and proposal contexts (Traum, 1994, p. 81)

[11] This notion has nothing to do with Wooldridge's computational grounding notion (Wooldridge, 2000).

5.5.3 Philosophy of language: Searle and Vanderveken

For Searle (Searle, 1983), performance of a speech act entails the expression of an Intentional state: for example the performance of an assertion entails automatically the expression of the speaker's belief about what he hase asserted. Even if an utterance was insincere an Intentional state has been *expressed*, and that state corresponds to a particular belief of the speaker in some way.

Vanderveken (Vanderveken, 1990; Vanderveken, 1991) has captured the subtle difference between *expressing* an Intentional state and *really being in* such a state by distinguishing *success conditions* from *non-defective performance conditions*, thus refining Searle's felicity conditions (Searle, 1969; Searle, 1979; Searle and Vanderveken, 1985). According to Vanderveken, when we assert p we *express* that we believe p (success condition), while the speaker's belief that p holds is a condition of non-defective performance (the sincerity condition).

Consider the 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 does not know 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 should be consistent).

5.5.4 Toward a new approach

In the sequel we will not consider the grounding process itself: just as Colombetti *et al.* or Chaib-draa *et al.* among aothers, we will also admit the simplifying hypothesis of perfect communication to avoid this problem. Thus we will consider only what is privately believed by participants and what is mutually believed (and thus public), *i.e.* which are the properties and features of the common ground, how agents can use it, deduce with it... and in particular how a proposed formula becomes accepted by both participants of the conversation. Thus Traum's schema presented above can be simplified by eliminating one step: when something has been proposed, it is mutually believed that it has been proposed. Indeed the shift from private to mutual belief state is immediate thanks to our public action hypotheses.

We will also concentrate our study on the link between the private and the public states. We will consider that this link is established by actions and particularly by speech acts in our case. In particular Speech Act Theory indicates what will become public after performance of speech acts.

We argue that the logic developed in Chapter 3 can be used to represent various states of Figure 5.3. In particular the private layer matches with our private beliefs. What has been proposed by the system or the user corresponds to the result of an utterance that has been perfectly understood by every agent. Indeed this corresponds to formulas of the form $G_{\{system,user\}}\, G_{system}\, \varphi$ (or $G_{\{system,user\}}\, G_{user}\, \varphi$). The shared state corresponds to a state where both participants have accepted the proposed proposition. It appears that this fits

the characterization of the group belief defended in the previous chapter. We can thus give a simplification of this figure, that can be formalized in our group belief logic (*cf.* Figure 5.4).

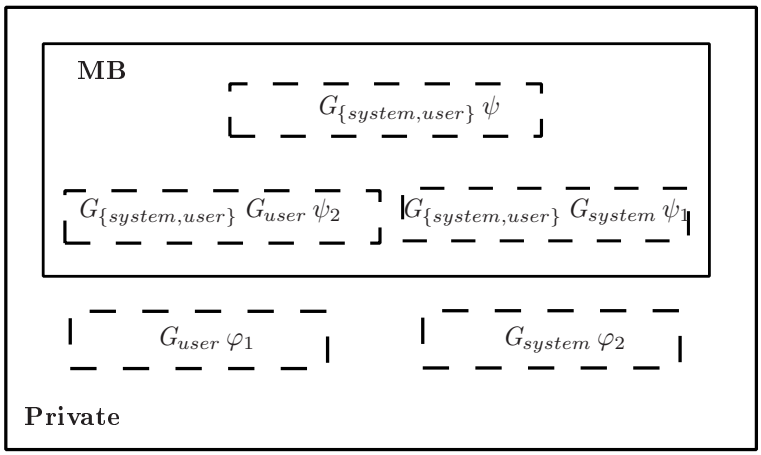

Figure 5.4: Simplification of Figure 5.3

In the sequel we will also refer to the G operator as the *groundedness* operator: it represents what has been grounded in the dialogue. We can remark that for Traum grounded information coincide with the set of what is mutually believed. Grounded information also corresponds to what is public for every agent overhearing the conversation (we will develop the extension from 2 participants to sets of overhearing agents below). Grounded propositions include as well what has been publicly expressed (or proposed) by an agent as what has been accepted by the participants of the conversation.

Justifications of the features. Our logic imposes that groundedness is a rational notion, *i.e.* a proposition and its contrary cannot be grounded in the same group. In particular we consider that an agent cannot assert something and its contrary. But this can be the case in front of two different groups, see the example below.

Grounded is a public[12] and thus objective notion: it refers to what can be observed, and only to that. It is different from other objective notions used to define ACLs such as that of social commitment of (Singh, 1998; Singh, 2000; Fornara and Colombetti, 2002; Verdicchio and Colombetti, 2003). To see the difference consider the request speech act where agent i asks agent j to pass him the salt. Thereafter it is established (given our hypothesis that the speech act is well and completely understood) that i has the intention that j passes him

[12]The public character of the groundedness is closely related to following simplifying hypothesis: we consider that every action is public, *i.e.* each agent perceives each action performed completely and soundly.

the salt and that nothing is grounded for j. In contrary, 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, whereas the requester is not committed to the fact that he performed a request.

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... or commitments.

A piece of information might be grounded in a group even when some agents privately disagree, as long as they do not publicly manifest their disagreement. Thus there is no link between the private and the public state of the conversation as illustrated by the following example:

EXAMPLE. *Consider three rational agents in a company. The agent 0 thinks privately for some reasons that his boss (agent 2) is smart. But this idea is not widespread in his department: agent 0 meets agent 1, a very charismatic agent who often claims publicly that his boss (agent 2) is dumb. They discuss about their boss and agent 0 asserts that he is really a moron (for some social reasons) and of course agent 1 confirms. At this moment the boss comes and enters the conversation. Soon he oriented the discussion on himself and agent 0 congratulates him by asserting he is smart. And agent 2 and agent 1 (given the boss' attendance) express their agreement.*

It is interesting to see that agent 0 expressed completely different points of view depending on the audience. As we are interested by what is public in a dialogue, we have to make precise which group of agents constitutes the audience because a speaker's behavior depends on who can hear what he says, as illustrated above. Thus we have to formalize such a dialogue by expliciting the different groups of hearers.

If we consider the above example, by asserting that his boss is a moron in front of agent 1, agent 0 has expressed that he believes that his boss is a moron (although 0 does not think so) and thus this proposition is grounded for agents 1 and 2. After acceptance by the second agent, it become grounded for the group of agents consisting of 0 and 1 that their boss is a moron. The contrary is grounded afterwards for the group of agents {0,1,2}.

In the sequel, we will not speak about piece of information grounded in general but rather about grounded for (or in front of) a group of agents. This extension allows to distinguish a private dialogue between two agents from a public debate. In the former a piece of information could be grounded between only two agents and stay secret for the other agents. In contrary in a public debate, every agent hearing the debate is aware of the grounded pieces of information.

5.5.5 Related Work: Boella, Nickles

The aim of the two following chapters (chapters 6 and 7) is to bridge the gap between the two main approaches by using our BDI-like logical framework ex-

tended with an operator formalizing what is public in the dialogue. Some other authors have studied ACLs with the same approaches. Among others we can cite the following.

5.5.5.1 Nickles *et al.*

Nickles *et al.* have proposed a logic of ostensible beliefs and intentions (Nickles, Fischer, and Weiss, 2005; Nickles, 2005). $Op(a_1, a_2, \varphi)$ reads "agent a_1 holds the ostensible belief A facing agent a_2". $OInt(a_1, a_2, \varphi)$ reads "agent a_1 facing agent a_2 exhibits the intention to make φ true". They only give a basic semantics to their logic, on top of which some principles are stated axiomatically. For example, their axiom (2) is: $Op(a_1, a_2, \varphi) \rightarrow \neg Op(a_1, a_2, \neg\varphi)$ (meaning that ostensible belief must stay consistent).

Their notion of ostensible mental states is very close to our notion of grounding mental states, and their operators can be translated into our logic. $Op(a_1, a_2, \varphi)$ corresponds to our $G_{\{a_1,a_2\}}\, G_{a_1}\, \varphi$, and $OInt(a_1, a_2, \varphi)$ to our $G_{\{a_1,a_2\}}\, Intend_{a_1}\, \varphi$.

They only give a basic semantics to their logic, on top of which some principles are stated axiomatically. For example, their axiom (2) becomes in our formalism $G_{\{a_1,a_2\}}\, G_{a_1}\, \varphi \rightarrow \neg G_{\{a_1,a_2\}}\, G_{a_1}\, \neg\varphi$. The latter is a theorem of our logic because the operator G_I satisfies the D-axiom for any group I. Similarly we can show that we can capture completely their notions of ostensible belief and intention with our semantically well-founded operator.

5.5.5.2 Boella *et al.*

Boella *et al.* (Boella et al., 2007) also propose to bridge the gap between mentalist and social approaches. They ascribe some roles to agents and associate directly to these roles beliefs and intentions, that are taken as public mental attitudes. They afterwards define ACL semantics thanks to these two operators: the authors give a formalization of both FIPA-ACL and Colombetti's commitment automaton.

Their operators have more or less the same meanings as Nickles' ones presented above. But instead of catching only the public feature of the dialogue, they also fit the beliefs ascribed to a role. Albeit this account appears attractive, it appears that it only shifts the problem of hypotheses on agents (such as sincerity or cooperation) from agents themselves to their role. In particular in the FIPA-ACL formalization, their account imposes roles to be sincere. No insincere or non-cooperative role can be ascribed to agents.

We use the expression "insincere role" in the following sense. Attitudes attributed to a role are public. Thus if an agent, in a given role, has a particular belief then this belief is public and thus known by every agents. In this sense, we argue that it is too strong. For example, the French president knows (/believes) the codes of the nuclear bombs. Thus as president, Sarkozy believes that codes are XXXXXX; but other French people cannot say that they believe that Sarkozy believes as president that codes are XXXXX.

5.6 Conclusion

In this chapter, we have presented an overview of the definition of Agent Communication Languages and in particular of the two main accounts to describe their semantics. We have argued that both accounts have their own benefits and drawbacks, and we have proposed a new approach to define their semantics based on the mental attitudes expressed *via* the performance of speech acts. We made the link with our group belief operator define in chapters 1 and 3.

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.

In the sequel we will show how we can apply this idea to formalize both kinds of semantics, *i.e.* mentalist semantics with FIPA-ACL and the application to the Contract Net Protocol (Chapter 6) and social semantics with the formalization of Colombetti's account (Chapter 7).

Chapter 6

Application: formalization of the mentalist approach

6.1 Introduction

In this chapter we will focus on two of the three separated modules composing the FIPA paradigm: the FIPA-ACL library and the set of Interactions Protocols (IPs). These two modules are devoted to manage and generate the dialog. The FIPA-ACL library exposes the set of allowed communicative primitives. It also gives their semantics in terms of a Feasibility Precondition (FP), that is the set of conditions necessary to allow the executability of the speech act, and a Rational Effect (RE) that is the effect intended by the locutor performing this act. It corresponds to the Perlocutionary Effect in speech act theory and is quite different from the Intentional Effect (the effect that is true as as soon as the act is performed) (*cf.* Section 5.2).

An Interaction Protocol describes speech acts sequences allowed in a dialogue. It ensures the coherence of the dialogue between agents. FIPA defines plenty of Interaction Protocols. In the sequel we will concentrate on the Contract Net Protocol (FIPA, 2002b).

Considering remarks made in the previous chapter about the need of a unified approach to ACLs semantics, we will give a new formalization of the FIPA-ACL semantics in terms of public facts formalized with our groundedness operator. In most ACLs based on mental attitudes and in particular in FIPA-ACL, only two kinds of speech acts are used: assertive and directive. Other illocutionary forces have raised only few interest[1]. We begin thus by defining those two primary speech acts following speech act theory principles (Section 6.2). We will apply this new semantics to formalize the Contract Net Protocol and exhibit some formal features (Section 6.3).

[1] Nevertheless there exists recent work trying to integrate declarative acts to FIPA (Demolombe and Louis, 2006).

6.2 Primitive FIPA speech acts

Following and extending the ACL syntax used in (FIPA, 2002a), a single communication act (CA) is denoted by $\langle i, J, K, \mathsf{ActName}, \varphi \rangle$, where i is the performing agent, J is a group of addressees (whereas FIPA allows only one). ActName is the name of the act (in our model not necessarily corresponding to exactly one speech act type, see below). The logical formula φ is the propositional content of the act. The parameter K, which is missing in FIPA, denotes a group of attending agents who overhear the respective utterance. We impose that $i \in K$, $J \subseteq K \setminus \{i\}$ and $J \neq \emptyset$. For a dialogue between only two agents i and j we have $J = \{j\}$ and $K = \{i, j\}$.

The distinction between the addressees J of a speech act and the group of agents K taking part in the conversation improves the usual FIPA-like characterization of speech acts: from the speech act theory standpoint, when a speaker talks to a subgroup J of K then the *success condition* (Searle, 1969; Vanderveken, 1990) applies only to J (but is evaluated from the point of view of the entire group K). Nevertheless, effects also obtain for the entire group K. This motivates that the addressees and the group must be distinguished, and must both be parameters of the speech act.

In the standard semantics of the FIPA CAs library (FIPA, 2002a) (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 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 effects. First, it is possible to desire and expect different kinds of RE for the same CA. Second, Searle shows in (Searle, 1969, 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. Third, 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) that we will provide in our model do not make any statement about mental attitudes, but specify the preconditions and effects in terms of formula grounded in the audience group K. 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 course of communication in terms of FPs and IEs—.

In the sequel we use the term *Social Attitudes Based Semantics* (SABS) for our modeling, 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.

6.2.1 Inform: Asserting information

6.2.1.1 Characterization

We start with the FIPA-S version of the semantics:

$\langle i, j, \mathsf{Inform}, \varphi \rangle$
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: $Bel_j\, \varphi$

$\langle i, j, \mathsf{Inform}, \varphi \rangle$ means: "agent i informs the receiver j that proposition φ is true". $U_j\varphi$ denotes that agent j is uncertain about φ, but thinks that φ is more likely than $\neg\varphi$. The essential preconditions of Inform in the FIPA-S semantics are thus that agent i truthfully believes what he asserts, and that the receiver does not have any definite opinion about the asserted proposition. The former condition is obviously unrealistic given a truly autonomous agent i. The latter is questionable. Indeed it disallows the use of Inform to convince the addressee. In (Gaudou et al., 2006) we have considered that the latter usage is crucial e.g. in the context of computational argumentation and argumentation-based negotiation and that the introduction of a conviction-act extending the syntax would not only be unnecessary and inelegant, but would also blur the fact that there exists a continuous 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 himself). In this dissertation, we stay close to the FIPA speech act library by distinguishing Inform and Confirm speech acts. A Confirm act informs the receiver in the case where he is uncertain and attempting to convince him.

We extend the Inform speech act by extending the individual addressee to groups J and introducing the group of overhearers K. Hence Inform acts take the form: $\langle i, J, K, \mathsf{Inform}, \varphi \rangle$. As FIPA preconditions are private mental attitudes, we do not keep them here. The preconditions of our actions are of two types: public relevance and public rationality. The *relevance precondition* of $\langle i, J, K, \mathsf{Inform}, \varphi \rangle$ is that i has not already expressed he believes φ, and the

same for J (that is: $\neg G_K\ G_i\ \varphi \wedge \neg G_K\ G_J\ \varphi$), and that J has not expressed that he does not believe φ (formally: $\neg G_K \neg G_J\ \varphi$ — otherwise the speech act would not be an inform act but a convince act). The *rationality precondition* corresponds to the fact that an agent must be publicly consistent, and means the agent has not expressed he does not believe φ (formally: $\neg G_K \neg G_i\ \varphi$).

The rational effect of Inform in FIPA-S is simply that the addressee believes what he has been told (in case the act succeeds). In FIPA-ACL, the *rational effect* (RE) roughly corresponds to the expected perlocutionary effect of the act. The RE is not directly added to the mental state of the addressee, but if this effect can be derived from the mental state (after the act performance) then the author of the act has achieved his aim. In fact, sincerity and credulity hypotheses always entail the rational effect.

But in fact, we can never guarantee such perlocutionary effects because we cannot control other agents' mental states.[2] Thus 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 either that the informing agent intends the adoption of a certain belief.

However, speech act theory says we cannot perform an action without **necessarily expressing** sincerity and preparatory conditions (Searle, 1983). The preparatory condition roughly corresponds to the relevance precondition of FIPA-ACL. Note that we adopt here a public point of view and do not impose that the agent is sincere and has checked the preparatory conditions. Usual BDI logics cannot capture this aspect of communication.

Thus, **expression** of such conditions is an **effect** of the act. When i informs J that φ, he expresses that he believes φ and that he intends J believes φ (formally: $G_K\ G_i\ \varphi \wedge G_K\ \mathit{Intend}_i\ G_J\ \varphi$), which is the expression of the sincerity condition.

One might think that it is too strong that only by performing a speech act an agent can ground a formula for the whole group. Moreover $G_K\ G_i\ \varphi$ can hold while neither agent i believes φ nor at least one agent of the group K believes that i believes φ. Such a situation could appear to be hypocrite. But in fact by asserting that φ, agent i commits himself in front of the whole group to his belief that φ. Thus formula $G_K\ G_i\ \varphi$ characterizes the acceptance by the group of the commitment. While members of the group can think privately that i has lied, they cannot deny that he has incurred a commitment. The acceptance is consequently implicit and immediate and does not require any discussion.[3] It only depends on our hypothesis of perfect communication.

The speaker also expresses the preparatory condition: he believes φ is not grounded for J yet (formally: $G_K\ G_i \neg G_J\ \varphi$).

These concerns lead to the following SABS semantics:

[2] From this point of view, Searle (Searle, 1969) shows that what we could name "perlocutionary act" cannot be a speech act (in the speech act theory sense), and was just a mistake of Austin (Austin, 1962).

[3] As we will see in the following chapter, an agent incurs a social commitment by performing an informing speech act in front of an attentive group of agents.

$\langle i, J, K, \mathsf{Inform}, \varphi \rangle$

FP: $\neg G_K\ G_i\ \varphi \wedge \neg G_K\ G_J\ \varphi \wedge \neg G_K \neg G_J\ \varphi \wedge \neg G_K \neg G_i\ \varphi$

IE: $G_K\ G_i\ \varphi \wedge G_K\ \mathit{Intend}_i\ G_J\ \varphi \wedge G_K\ G_i \neg G_J\ \varphi$

As usual we define $\langle i, J, K, \mathsf{InformIf}, \varphi \rangle$ as an abbreviation:

$$\langle i, J, K, \mathsf{InformIf}, \varphi \rangle \stackrel{def}{=} \langle i, J, K, \mathsf{Inform}, \varphi \rangle \cup \langle i, J, K, \mathsf{Inform}, \neg\varphi \rangle$$

Hence:

$$\mathit{Done}(\langle i, J, K, \mathsf{InformIf}, \varphi \rangle) \equiv$$
$$\mathit{Done}(\langle i, J, K, \mathsf{Inform}, \varphi \rangle) \vee \mathit{Done}(\langle i, J, K, \mathsf{Inform}, \neg\varphi \rangle)$$

6.2.1.2 Properties

In this section we show that the above characterization of the Inform speech act gives interesting formal results. In particular it forbids the performance of some syntactically well-formed speech acts which could appear counter-intuitive or too strong.

What about inform-actions with some belief of super groups as propositional content? Can an agent make a group believe something in this way? We have the following theorem:

THEOREM. *The action* $\langle i, J, K, \mathsf{Inform}, G_{K'}\ \varphi \rangle$ *is inexecutable, for each* K' *such that* $K \subseteq K' \subseteq AGT$.

PROOF. *We will prove that the preconditions of this act are inconsistent. Let* K' *be a supergroup of* K, i.e. $K \subseteq K' \subseteq AGT$. *In particular we have:*

$$\vdash \mathit{FP}(\langle i, J, K, \mathsf{Inform}, G_{K'}\ \varphi \rangle) \rightarrow \neg G_K \neg G_i\ G_{K'}\ \varphi \wedge \neg G_K\ G_i\ G_{K'}\ \varphi$$

From Theorems (3.2) and (3.3) of Section 3.3.1 Chapter 3, we can prove the equivalences, for $i \in K$ *and* $K \subseteq K'$*:*

$$\vdash \neg G_K \neg G_i\ G_{K'}\ \varphi \leftrightarrow \neg G_K \neg G_{K'}\ \varphi$$

$$\vdash \neg G_K \neg G_{K'}\ \varphi \leftrightarrow G_{K'}\ \varphi$$

Similarly from theorem (3.2), we can show that:

$$\vdash \neg G_K\ G_i\ G_{K'}\ \varphi \leftrightarrow \neg G_K\ G_{K'}\ \varphi$$

$$\vdash \neg G_K\ G_{K'}\ \varphi \leftrightarrow \neg G_{K'}\ \varphi$$

The precondition of the action $\langle i, J, K, \mathsf{Inform}, G_{K'}\ \varphi \rangle$ *is inconsistent and thus this kind of acts is inexecutable.* □

If an agent could perform the act $\langle i, J, K, \mathsf{Inform}, G_{K'}\ \varphi \rangle$, one of its effects would be $G_K\ G_i\ G_{K'}\ \varphi$, which is equivalent to $G_{K'}\ \varphi$. This theorem highlights

an important property of our logic: if an agent could perform such a speech act he could ground a formula φ for the whole group without possible discussion.

Moreover this theorem sheds a new light on the seemingly too powerful theorem (3.2) ($G_I\,\varphi \leftrightarrow G_{I'}\,G_I\,\varphi$) and its counterpart (3.3). In particular, the implication $G_{I'}\,G_I\,\varphi \rightarrow G_I\,\varphi$ says that when it is grounded for a subgroup I' of I that $G_I\,\varphi$ then *de facto* it is grounded for I that φ, and seems to give too much power to a subgroup. But the above theorem expresses that no agent of I' can express a formula in the scope of operator G_I, *i.e.* he cannot establish formulas such as $G_{I'}\,G_I\,\varphi$. Thus such a formula can only hold if $G_I\,\varphi$ holds, which is a quite intuitive property.

THEOREM. *The following actions are inexecutable:*

$$\langle i, J, K, \mathsf{Inform}, \mathit{Done}_{k:\alpha}\,\top\rangle,\ \forall k \in K \tag{6.1}$$

$$\langle i, J, K, \mathsf{Inform}, \neg\mathit{Done}_{k:\alpha}\,\top\rangle,\ \forall k \in K \tag{6.2}$$

$$\langle i, J, K, \mathsf{Inform}, \varphi \wedge \neg G_i\,\varphi\rangle \tag{6.3}$$

PROOF. *Theorems (6.1) and (6.2) can be proven immediately from the above theorem thanks to axioms ($PA_{I,\alpha}$) and ($NA_{I,\alpha}$) of Section 3.3.7 of Chapter 3:*

$$G_I\,\mathit{Done}_\alpha\,\top \leftrightarrow \mathit{Done}_\alpha\,\top$$

$$G_I\,\neg\mathit{Done}_\alpha\,\top \leftrightarrow \neg\mathit{Done}_\alpha\,\top$$

For theorem (6.3) we will prove that the preconditions of this act are inconsistent. We have:

1. $\vdash \boldsymbol{Prec}(\langle i, J, K, \mathsf{Inform}, \varphi \wedge \neg G_i\,\varphi\rangle) \rightarrow \neg G_K\,\neg G_i\,(\varphi \wedge \neg G_i\,\varphi)$
2. $\vdash G_i\,(\varphi \wedge \neg G_i\,\varphi) \leftrightarrow G_i\,\varphi \wedge G_i\,\neg G_i\,\varphi$, *theorem of any normal modal logic*
3. $\vdash G_i\,\varphi \wedge G_i\,\neg G_i\,\varphi \leftrightarrow G_i\,\varphi \wedge \neg G_i\,\varphi$, *Axiom* ($5_{G_I}$)
4. $\vdash \neg G_i\,(\varphi \wedge \neg G_i\,\varphi) \leftrightarrow \top$, *from 2., 3. and LP*
5. $\vdash G_K\,\neg G_i\,(\varphi \wedge \neg G_i\,\varphi) \leftrightarrow G_K\,\top$, *from 5. by* RN_{G_I}
6. $\vdash \neg G_K\,\neg G_i\,(\varphi \wedge \neg G_i\,\varphi) \leftrightarrow \neg G_K\,\top$, *from 6. and LP*
7. $\vdash \neg G_K\,\top \leftrightarrow \bot$, *theorem of any normal modal logic*
8. $\vdash \boldsymbol{Prec}(\langle i, J, K, \mathsf{Inform}, \varphi \wedge \neg G_i\,\varphi\rangle) \rightarrow \bot$, *from 1., 7. and LP*
9. $\vdash \neg\boldsymbol{Prec}(\langle i, J, K, \mathsf{Inform}, \varphi \wedge \neg G_i\,\varphi\rangle)$, *from 8. and LP*

□

6.2.2 Request: Requesting an action to be done

Again, we state the FIPA version of the semantics first:

$\langle i, j, \mathsf{Request}, \alpha \rangle$
FP: $FP(\alpha)[i\backslash j] \wedge Bel_i\, Agent(j, \alpha) \wedge Bel_i \neg Intend_j\, Done(\alpha)$
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 $Intend_j\, Done(\alpha)$ denotes that j has the intention (Sadek, 1992) that $Done(\alpha)$ holds (*i.e.* that action α has just been performed successfully).

Obviously, these specifications again make 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 himself.

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). In our formalism $(i{:}\alpha)$ denotes that i is the author of action α and makes superfluous the FIPA *Agent* predicate.

We extend the Request speech act by introducing the group K of overhearers: $\langle i, J, K, \mathsf{Request}, \varphi \rangle$ means "agent i requests a subset J of group of agents K to perform some action having φ as effect, K attending". The *relevance precondition* is: it is not grounded for K that

1. i intends that φ,
2. J intends that φ, and
3. J does not intend φ (otherwise the act would be close to a persuasion speech act).

The *rationality precondition* is that it is not grounded for K that i does not intend φ.

The FIPA RE ($Done\,(\alpha)$) just specifies that the intended perlocutionary effect of this communication act is to get α done. Of course this effect is too strong, and for us the *effects* are:

1. i intends that φ (expression of the *sincerity condition*), and
2. i expresses he believes that J does not intend that φ be true (expression of the *preparatory condition*).

We did not define what group intention is. Here, we only consider individual actions, whose authors are individual agents which do not need other agents (versus group actions, group intention, teamwork... as *e.g.* studied in (Cohen and Levesque, 1994)). We thus have two kinds of request RequestEach and RequestSome (whereas there is only one in FIPA), depending on the meaning we give to "J intends to perform action α". Furthermore, due to our definition of intention, here the negation of a choice is more appropriate than a negation of an intention[4]. Thus we have the two following formalizations:

$$\langle i, J, K, \mathsf{RequestSome}, J{:}\alpha \rangle \stackrel{def}{=} \langle i, J, K, \mathsf{RequestSome}, \bigvee_{j \in J} Done((j{:}\alpha)) \rangle$$

$$\text{FP: } \neg G_K\, Intend_i \bigvee_{j \in J} Done((j{:}\alpha)) \wedge \left(\neg G_K \bigvee_{j \in J} Intend_j\, Done((j{:}\alpha)) \right)$$
$$\wedge \neg G_K \left(\bigwedge_{j \in J} \neg Choice_j\, Done((j{:}\alpha)) \right) \wedge \neg G_K \bigwedge_{j \in J} \neg Choice_i\, Done((j{:}\alpha))$$

$$\text{IE: } G_K\, Intend_i \left(\bigvee_{j \in J} Done((j{:}\alpha)) \right) \wedge G_K\, G_i \left(\bigwedge_{j \in J} \neg Choice_j\, Done((j{:}\alpha)) \right)$$

which specifies that i intends that at least one agent among J performs the requested action α.

Second, we define:

$$\langle i, J, K, \mathsf{RequestEach}, J{:}\alpha \rangle \stackrel{def}{=} \langle i, J, K, \mathsf{RequestEach}, \bigwedge_{j \in J} Done((j{:}\alpha)) \rangle$$

$$\text{FP: } \neg G_K\, Intend_i \bigwedge_{j \in J} Done((j{:}\alpha)) \wedge \left(\neg G_K \bigwedge_{j \in J} Intend_j\, Done((j{:}\alpha)) \right)$$
$$\wedge \neg G_K \left(\bigvee_{j \in J} \neg Choice_j\, Done((j{:}\alpha)) \right) \wedge \neg G_K \bigvee_{j \in J} \neg Choice_i\, Done((j{:}\alpha))$$

$$\text{IE: } G_K\, Intend_i \left(\bigwedge_{j \in J} Done((j{:}\alpha)) \right) \wedge G_K\, G_i \left(\bigvee_{j \in J} \neg Choice_j\, Done((j{:}\alpha)) \right)$$

which specifies that i intends that each agent of J perform the requested action α.

[4] Indeed by definition of the intention, the negation of an intention that φ holds is equivalent to the fact that either the agent believes that φ holds, or the agent believes that the action will be performed independently of any of his actions, or the agent does not choose that φ holds anymore. Only the last is relevant here.

For compatibility reasons, we choose the RequestSome speech act as the "official" request act and define:

$$\langle i, J, K, \mathsf{Request}, \alpha \rangle \stackrel{def}{=} \langle i, J, K, \mathsf{RequestSome}, J{:}\alpha \rangle$$

6.2.3 Confirm and Disconfirm

FIPA also defines the acts Confirm (for the confirmation of an uncertain information) and its counterpart Disconfirm (the utterance that φ is false whereas the hearer believes that it holds) as primitives. Confirm and Disconfirm are close to Inform assertive speech acts, but they need stronger preconditions to be performed. As we do not have Sadek's operator of uncertainty, we consider that a Confirm and Disconfirm speech act can be performed only if the hearers have expressed that they envisage φ. Moreover we impose for the Confirm (resp. Disconfirm) that the speaker has not yet uttered that φ (resp. $\neg\varphi$) holds (*relevance preconditions*). It is a bit stronger than the FIPA version of the Confirm (resp. Disconfirm) and also than the version in (Searle, 1969). We keep also the *rationality precondition* that is that i has not already expressed φ (resp. $\neg\varphi$).

Moreover we stay close to the Inform speech act concerning effects: after a Confirm it is grounded that the speaker believes what he has uttered, that he has the intention that the group of addresses believes φ (expression of sincerity conditions) and that he believes they envisage that φ holds (expression of preparatory condition).

We thus have the formalization:

$\langle i, J, K, \mathsf{Confirm}, \varphi \rangle$
FP: $G_K \neg G_J \neg\varphi \wedge \neg G_K\, G_i\, \varphi \wedge \neg G_K \neg G_i\, \varphi$
IE: $G_K\, G_i\, \varphi \wedge G_K\, \mathit{Intend}_i\, G_J\, \varphi \wedge G_K\, G_i \neg G_J \neg\varphi$

Concerning the Disconfirm speech act, the effects are that it is grounded that the speaker believes that φ does not hold, that he has the intention that the group of addressees believes $\neg\varphi$ (expression of sincerity conditions) and that he believes they envisage that φ holds (expression of preparatory condition).

$\langle i, J, K, \mathsf{Disconfirm}, \varphi \rangle$
FP: $G_K \neg G_J \neg\varphi \wedge \neg G_K\, G_i \neg\varphi \wedge \neg G_K \neg G_i \neg\varphi$
IE: $G_K\, G_i \neg\varphi \wedge G_K\, \mathit{Intend}_i\, G_J \neg\varphi \wedge G_K\, G_i \neg G_J \neg\varphi$

Thanks to the above definitions, we can now translate each speech act of the FIPA-ACL library in our framework. We begin by giving a case study to illustrate the new semantics of these speech acts. Then we show how to translate speech acts of FIPA, afterwards we express additional preconditions and effects for dialogue acts related to Interaction Protocols.

6.2.4 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 nature, any other BDI-based ACL semantics:

- Pre- and post-conditions of communication acts only depend on publicly observable agent behavior, thus being fully verifiable.
- It is possible to perform communication acts whose content is inconsistent with the beliefs and intentions of the participating agents.
- Communication acts may address groups of agents.
- Multiple communication acts may be uttered by the same sender, but with mutually inconsistent contents (even towards nested groups);

In addition, the example will show 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 6.1, we consider this to be simpler and formally clearer compared to the use of social commitments in the sense of (Singh, 2000).

The interaction roughly follows protocols for *purchase negotiation dialogue games* as known from, *e.g.* (McBurney et al., 2003), 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.

Some of the agents in our scenario are insincere, asserting information they do not believe. As we have seen, SABS allows such agents to interact without getting into conflict with the norms of the ACL semantics. This is not in order to encourage agents to lie, but rather to neatly separate the levels of private beliefs and intentions on the one hand, and publicly expressed opinions and ostensible intentions on the other. We explicitly take into account that agents might be insincere, and respond by giving the possibility of modeling the difference between private and expressed beliefs and intentions. Thus, although no imposed ACL semantics can ever guarantee a certain agent behavior (like that an agent sticks to her publicly uttered beliefs over time), SABS hinders the agents to give false account of their public attitudes. Of course, the (private) beliefs and intentions of the participants are usually not available to an external

observer like the protocol designer of an open multiagent system, but are given for the agents below for explanatory purposes.

Our scenario consists of three agents $AGT = \{s_1, b_1, b_2\}$, representing potential car seller 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:

Utterance no. *sender→receiver:* **Descriptive act title**
Message [5]
Effect

Note that the interaction course consists of multiple interlaced conversations among different sender/receiver pairs and different overhearers (*i.e.* different audiences). In particular, agent b_2 is involved in two selling dialogues at the same time. *Effect* is optional and gives the effect of the act in terms of grounded formulas, according to SABS.

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. Note that utterances **U1** to **U5** are meta-dialogical: their aim is to initialize interaction by managing negotiation between agents to dialog with each others. Semantics of these acts is thus not detailed here.

U1 $s_1 \rightarrow \{b_1, b_2\}$**: Initialize dialogue**
$\langle s_1, \{b_1, b_2\}, \{s_1, b_1, b_2\}, \mathsf{RequestEach}, enterDialogue(\theta_1)\rangle$

U2 $b_1 \rightarrow \{s_1\}$**: Enter dialogue**
$\langle b_1, \{s_1\}, \{s_1, b_1, b_2\}, \mathsf{Agree}, enterDialogue(\theta_1)\rangle$

U3 $b_2 \rightarrow \{s_1\}$**: Enter dialogue**
$\langle b_2, \{s_1\}, \{s_1, b_1, b_2\}, \mathsf{Agree}, enterDialogue(\theta_1)\rangle$

U4 $s_1 \rightarrow \{b_2\}$**: Initialize dialogue**
$\langle s_1, \{b_2\}, \{s_1, b_2\}, \mathsf{Request}, enterDialogue(\theta_2)\rangle$

U5 $b_2 \rightarrow \{s_1\}$**: Enter dialogue**
$\langle b_2, \{s_1\}, \{s_1, b_2\}, \mathsf{Agree}, enterDialogue(\theta_2)\rangle$

PI_{s_1}**:** $Bel_{s_1}\, discounts$

U6 $s_1 \rightarrow \{b_1, b_2\}$**: Information about discount**
$\langle s_1, \{b_1, b_2\}, \{s_1, b_1, b_2\}, \mathsf{Inform}, \neg discounts\rangle$
Effect:
$G_{\{s_1,b_1,b_2\}}\, G_{s_1}\, \neg discounts \wedge G_{\{s_1,b_1,b_2\}}\, Intend_{s_1}\, G_{\{b_1,b_2\}}\, \neg discounts$
$\wedge G_{\{s_1,b_1,b_2\}}\, G_{s_1}\, \neg G_{\{b_1,b_2\}}\, \neg discouts$

[5] Using syntactical macros according to (FIPA, 2002a). Only in case the message primitives are semantically relevant in our context, the respective macros are expanded.

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 can be given, but that might reduce the seller's individual profit).

U7 $s_1 \rightarrow \{b_2\}$**: Information about discount**
$\langle s_1, \{b_2\}, \{s_1, b_2\}, \mathsf{Inform}, discounts\rangle$
Effect:
$G_{\{s_1,b_2\}}\, G_{s_1}\, discounts \wedge G_{\{s_1,b_2\}}\, Intend_{s_1}\, G_{b_2}\, discount$
$\wedge G_{\{s_1,b_2\}}\, G_{s_1}\, \neg G_{b_2}\, discounts$
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 valuable customer whereas b_1 is not).

U8 $b_2 \rightarrow \{s_1\}$**: Query if car type has high accident rate**
$\langle b_2, \{s_1\}, \{s_1, b_2\}, \mathsf{Request}, InformIfAccidentRateHigh\rangle$
Effect:
$G_{\{s_1,b_2\}}\, Intend_{b_2}\, Done(s_1 : InformIfAccidentRateHigh) \wedge$
$G_{\{s_1,b_2\}}\, \neg G_{b_2}\, Int_{s_1} Done(s_1 : InformIfAccidentRateHigh)$, with
$InformIfAccidentRateHigh \stackrel{def}{=} \langle s_1, \{b_2\}, \{s_1, b_2\}, \mathsf{InformIf}, accidentRateHigh(\theta)\rangle$

$PI_{s_1} : Bel_{s_1}\, accidentRateHigh(\theta_1)$

U9 $s_1 \rightarrow \{b_2\}$**: Information about accident rate**
$\langle s_1, \{b_2\}, \{s_1, b_2\}, \mathsf{Inform}, \neg accidentRateHigh(\theta)\rangle$
Effect:
$G_{\{s_1,b_2\}}\, G_{s_1}\, \neg accidentRateHigh(\theta) \wedge G_{\{s_1,b_2\}}\, G_{b_2} \neg accidentRateHigh(\theta)$
$\wedge G_{\{s_1,b_2\}}\, G_{s_1}\, \neg G_{b_2}\, \neg accidentRateHigh(\theta)$
Seller s_1 asserted $\neg accidentRateHigh(\theta_1)$ while privately believing the opposite. Note that at this point, b_2 can adopt privately the belief that $\neg accidentRateHigh(\theta_1)$ or not regarding how he trusts s_1, regardless the fact that $G_{\{s_1,b_2\}}\, G_{s_1}\, \neg accidentRateHigh(\theta)$ holds.

U10 $b_2 \rightarrow \{s_1\}$**: Propose to buy at a certain price**
$\langle b_2, \{s_1\}, \{s_1, b_2\}, \mathsf{Propose}, buy(\theta_2, 10000£)\rangle$

U11 $s_1 \rightarrow \{b_2\}$**: Accept proposal**
$\langle s_1, \{b_2\}, \{s_1, b_2\}, \mathsf{AcceptProposal}, buy(\theta_2, 10000£)\rangle$
Effect (with the previous act):
$G_{\{s_1,b_2\}}\, Intend_{b_2}\, buy(\theta_2, 10000£)$ (*i.e.* b_2 is publicly committed to buy θ_2 at the price of $10000£ now).

In this case study we only highlight the use of our new semantics and in particular the link that the private layer has with the public one. Note that we did not consider here the whole sequence of utterances: this issue is the object of the following section.

We have redefined the FIPA semantics thanks to the grounded operator and thus defined it in terms of public mental attitudes in particular for the

four primitives of the FIPA speech act library. We now apply this semantics to formalize Interaction Protocols and more particulary the Contract Net Protocol.

6.3 Application to the Contract Net Protocol

6.3.1 Description of the CNP protocol

We revisit here the well-known Contract Net Protocol (CNP) (FIPA, 2002a). Following this protocol, an agent (the *initiator*) wants a task to be performed by one or more agents and wishes some optimization of this performance by giving himself the choice between several offers of action performance. The protocol flaw is described in the figure 6.1.

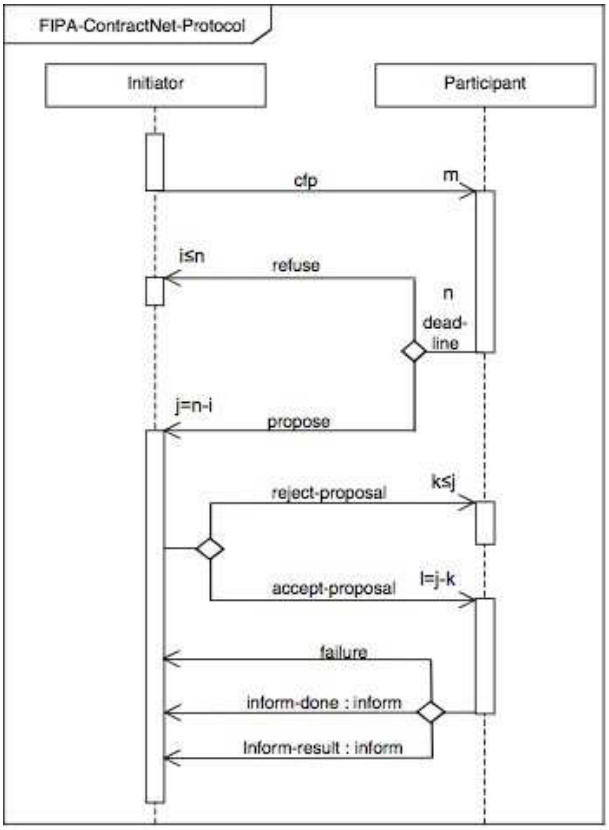

Figure 6.1: Contract Net Protocol (FIPA, 2002b)

In this protocol, there is only one *initiator* and one and more *participants*. The initiator begins the dialogue by sending a Call for Proposal (Cfp) to all the participants. Before the end of a certain deadline, they have to respond

to him by a proposition (Propose) or by a refusal (Refuse). The initiator has to choose between the propositions to accept some (AcceptProposal) and to refuse explicitly the others (RefuseProposal). The chosen ones must inform him about the result: Failure or InformDone, which ends the dialogue. At any time, the initiator can interrupt the dialogue by canceling it (Cancel). To be more precise, the FIPA protocol forecasts the case that there can be problems of understanding between the two agents. If an agent doesn't understand a message he can perform a NotUnderstood speech act in the aim to inform other agents of its misunderstanding. In this case, the dialogue is canceled.

Our aim is to formalize speech acts used in the CNP by mens of the above primitives. Then we show the soundness and completeness of our system with respect to the original CNP.

6.3.2 Speech acts formalization

6.3.2.1 Simplifications.

Due to our logical language (propositional modal logic), we need to do some simplifications in the translation of original definitions of FIPA. In particular, the FIPA Content Language (FIPA, 2002c) uses referential expressions and expressions with existential and universal quantifiers. We have to to simplify such expressions. Of course we lose lot of expressivity and this limitation is very severe for actual systems. But we argue that this simplification does not disturb the correct execution of the dialogue. We only adapt the definition of the speech acts. The use of a more expressive Content Language would be straightforward. Moreover we do not have the notion of deadline in our language, so we consider that every agent answers to the cfp. Those that would not respect the deadline are considered as if they have done a refuse act. Each simplification will be justified in the sequel on a case-by-case basis.

Before the next paragraph, we stress that these simplifications do not change the spirit of the protocol. Their role is to highlight the benefit of our work for the definition of a new semantics without (for our purpose) useless complex details.

6.3.2.2 Call for Proposal ($\langle i, J, K, \mathsf{Cfp}, J{:}\alpha\rangle$)

In FIPA Communicative Acts library, the call for proposal is viewed as a request to inform the speaker if the hearer has the intention to perform some action with a referential precondition. As in (Paurobally, Cunningham, and Jennings, 2005) and due to limitations on the Content Language (in particular we do not have the notion of referents), we consider the call for proposal act as a kind of request to propose to perform a given action (without any additional referential precondition).

Thus we assume that a call for proposal is nothing else that a request where the requested action is a proposal to perform some action α, *i.e.* $\langle j, \mathsf{Propose}, i, j{:}\alpha\rangle$. Finally the precondition and effect of a call for proposal are:

$$\langle i, J, K, \mathsf{Cfp}, \alpha\rangle$$

$$\text{FP: } \neg G_K\, \mathit{Intend}_i \bigvee_{j\in J} \mathit{Done}_{\langle j,i,K,\mathsf{Propose},j:\alpha\rangle}\top$$

$$\wedge \neg G_K \bigvee_{j\in J} \mathit{Intend}_j\, \mathit{Done}_{\langle j,i,K,\mathsf{Propose},j:\alpha\rangle}\top$$

$$\wedge \neg G_K \bigwedge_{j\in J} \neg \mathit{Choice}_j\, \mathit{Done}_{\langle j,i,K,\mathsf{Propose},j:\alpha\rangle}\top$$

$$\wedge \neg G_K \bigwedge_{j\in J} \neg \mathit{Choice}_i\, \mathit{Done}_{\langle j,i,K,\mathsf{Propose},j:\alpha\rangle}\top$$

$$\text{IE: } G_K\, \mathit{Intend}_i \bigvee_{j\in J} \mathit{Done}_{\langle j,i,K,\mathsf{Propose},j:\alpha\rangle}\top$$

$$\wedge G_K\, G_i \bigwedge_{j\in J} \neg \mathit{Choice}_j\, \mathit{Done}_{\langle j,i,K,\mathsf{Propose},j:\alpha\rangle}\top$$

6.3.2.3 Propose ($\langle i, J, K, \mathsf{Propose}, i{:}\alpha\rangle$)

The Propose speech act means that the agent offers to perform a certain action. For FIPA-ACL it can be reduced to an inform act that the speaker will adopt the intention to perform the action if the hearer intends that he does it. Of course in our formalism we impose that the left part of this implication is public: by proposing to do α, the agent i informs the group J (and thus expresses publicly) that he will adopt the intention to perform the action in the case where they express their intention that he performs it. Once again this action does not imply anything about any private intention adoption.

$$\langle i, J, K, \mathsf{Propose}, i{:}\alpha\rangle \stackrel{def}{=} \langle i, J, K, \mathsf{Inform}, \left(G_K \bigvee_{j\in J} \mathit{Intend}_j\, i{:}\alpha \right) \rightarrow \mathit{Intend}_i\, i{:}\alpha\rangle$$

This is a very general definition of the Propose act because it allows spontaneous propositions. It is relevant in an actual dialog between human beings but it appears a bit too weak in an interaction protocol. Thus we strengthen it in the sequel to avoid spontaneous propositions that could affect Interaction Protocols properties (*e.g.* termination). Thus we consider that a proposition can only be made after having been requested (for example after a call of proposals), and we strengthen the preconditions of Inform in consequence: $G_K\, \mathit{Intend}_j\, \langle i, J, K, \mathsf{Propose}, i{:}\alpha\rangle$.

Note that this property is not CNP dependant. This relevance feature should be imposed in the general case. We could moreover define a spontaneous propose act but a protocol using such a speech act would not have termination property. In human-human dialogs, propositions can of course be spontaneous, but they nevertheless need some kind of coherence with the current dialogue. Here we only impose a kind of strong coherence.

Thus its semantics follows from the above remarks:

$$\begin{aligned}
&\langle i, J, K, \mathsf{Propose}, i{:}\alpha\rangle \\
&\quad \text{FP: } \neg G_K\, G_i\, \psi \wedge \neg G_K\, G_J\, \psi \wedge \neg G_K\, \neg G_J\, \psi \wedge \neg G_K\, \neg G_i\, \psi \\
&\qquad \wedge G_K \bigvee_{j \in J} \mathit{Intend}_j\, \langle i, J, K, \mathsf{Propose}, i{:}\alpha\rangle \\
&\quad \text{IE: } G_K\, G_i\, \psi \wedge G_K\, \mathit{Intend}_i\, G_J\, \psi \wedge G_K\, G_i\, \neg G_J\, \psi
\end{aligned}$$

where $\psi = \left(G_K \bigvee_{j \in J} \mathit{Intend}_j\, i{:}\alpha\right) \rightarrow \mathit{Intend}_i\, i{:}\alpha$

Note that the first part of the IE can be reduced to:

$$G_K \bigvee_{j \in J} \mathit{Intend}_j\, i{:}\alpha \rightarrow G_K\, \mathit{Intend}_i\, i{:}\alpha$$

meaning that as soon as an agent of the group J has expressed his intention that i performs the action α (with a request or an acceptance of the proposal), agent i adopts publicly the intention to perform it.

6.3.2.4 Refuse ($\langle i, J, K, \mathsf{Refuse}, i{:}\alpha\rangle$)

Agents can refuse to perform some action, for any reason. To simplify, we do not consider here this reason. The Refuse speech act has an Inform part: it informs that the agent does not have the intention to perform the action (*i.e.* $\neg \mathit{Choice}_i\, \mathit{Done}\, (i{:}\alpha)$). But it cannot be reduced just to an inform. Before, the agent must have received a request to do something in order to perform a refusal. So the semantics of Refuse is the combination of the one of the Inform, with the precondition that an agent requested him to perform the action:

$$\begin{aligned}
&\langle i, J, K, \mathsf{Refuse}, i{:}\alpha\rangle \\
&\quad \text{FP: } \neg G_K\, G_i\, \neg \mathit{Choice}_i\, \mathit{Done}\, (i{:}\alpha) \wedge \neg G_K\, G_J\, \neg \mathit{Choice}_i\, \mathit{Done}\, (i{:}\alpha) \\
&\qquad \wedge \neg G_K\, \neg G_J\, \neg \mathit{Choice}_i\, \mathit{Done}\, (i{:}\alpha) \wedge \neg G_K\, \neg G_i\, \neg \mathit{Choice}_i\, \mathit{Done}\, (i{:}\alpha) \\
&\qquad \wedge G_K\, \mathit{Intend}_j\, \mathit{Done}\, (i{:}\alpha) \\
&\quad \text{IE: } G_K\, G_i\, \neg \mathit{Choice}_i\, \mathit{Done}\, (i{:}\alpha) \\
&\qquad \wedge G_K\, \mathit{Intend}_i\, G_J\, \neg \mathit{Choice}_i\, \mathit{Done}\, (i{:}\alpha) \\
&\qquad \wedge G_K\, G_i\, \neg G_J\, \neg \mathit{Choice}_i\, \mathit{Done}\, (i{:}\alpha)
\end{aligned}$$

In the case of the Contract Net Protocol, agents have the choice between proposing to perform the action α and refusing to propose, *i.e.* performing the act: $\langle i, j, K, \mathsf{Refuse}, i{:}\alpha\rangle$, where j is the initiator.

6.3.2.5 Accept and Reject Proposal

Accept and Reject Proposalacts are very similar. By the AcceptProposal (respectively RejectProposal) act, the speaker accepts (respectively refuses) the proposition made to him, *i.e.* he informs the hearer that he has (respectively

does not have) the intention that the addressee performs the action he proposed. FIPA-ACL defines them simply as Inform:

$$\langle i, J, K, \mathsf{AcceptProposal}, j{:}\alpha\rangle \stackrel{def}{=} \langle i, J, K, \mathsf{Inform}, \mathit{Intend}_i\ \mathit{Done}\ (j{:}\alpha)\rangle$$
$$\langle i, J, K, \mathsf{RejectProposal}, j{:}\alpha\rangle \stackrel{def}{=} \langle i, J, K, \mathsf{Inform}, \neg \mathit{Choice}_i\ \mathit{Done}\ (j{:}\alpha)\rangle$$

This definition seems too weak: it does not consider that AcceptProposal (resp. RejectProposal) can be performed only after a proposal. We must then modify the semantic of FIPA-ACL by strengthening the preconditions to completly capture the notion of acceptance and refusal of a proposal. Moreover in our framework, actions are performed only by a single agent. As these two acts are addressed to agents having proposed α, we lighten the notation by considering only the case where J is reduced to one agent j:

$\langle i, j, K, \mathsf{AcceptProposal}, j{:}\alpha\rangle$

FP: $\neg G_K\ G_i\ \mathit{Intend}_i\ \mathit{Done}\ (j{:}\alpha) \wedge \neg G_K\ G_J\ \mathit{Intend}_i\ \mathit{Done}\ (j{:}\alpha)$
$\wedge \neg G_K \neg G_J\ \mathit{Intend}_i\ \mathit{Done}\ (j{:}\alpha) \wedge \neg G_K \neg G_i\ \mathit{Intend}_i\ \mathit{Done}\ (j{:}\alpha)$
$\wedge\ G_K\ \mathit{Intend}_j\ G_i\ (G_K\ \mathit{Intend}_i\ j{:}\alpha \rightarrow \mathit{Intend}_j\ j{:}\alpha)$

IE: $G_K\ G_i\ \mathit{Intend}_i\ \mathit{Done}\ (j{:}\alpha) \wedge G_K\ \mathit{Intend}_i\ G_J\ \mathit{Intend}_i\ \mathit{Done}\ (j{:}\alpha)$
$\wedge\ G_K\ G_i \neg G_J\ \mathit{Intend}_i\ \mathit{Done}\ (j{:}\alpha)$

Remark: The effect can be simplified because $G_K\ G_i\ \mathit{Intend}_i\ j{:}\alpha$ $\leftrightarrow\ G_K\ \mathit{Intend}_i\ j{:}\alpha$ is valid in our groundedness logic presented in Chapter 3.

$\langle i, j, K, \mathsf{RejectProposal}, j{:}\alpha\rangle$

FP: $\neg G_K\ G_i \neg \mathit{Choice}_i\ \mathit{Done}\ (j{:}\alpha) \wedge \neg G_K\ G_J \neg \mathit{Choice}_i\ \mathit{Done}\ (j{:}\alpha)$
$\wedge \neg G_K \neg G_J \neg \mathit{Choice}_i\ \mathit{Done}\ (j{:}\alpha) \wedge \neg G_K \neg G_i \neg \mathit{Choice}_i\ \mathit{Done}\ (j{:}\alpha)$
$\wedge\ G_K\ \mathit{Intend}_j\ G_i\ (G_K\ \mathit{Intend}_i\ j{:}\alpha \rightarrow \mathit{Intend}_j\ j{:}\alpha)$

IE: $G_K\ G_i \neg \mathit{Choice}_i\ \mathit{Done}\ (j{:}\alpha) \wedge G_K\ \mathit{Intend}_i\ G_J \neg \mathit{Choice}_i\ \mathit{Done}\ (j{:}\alpha)$
$\wedge\ G_K\ G_i \neg G_J \neg \mathit{Choice}_i\ \mathit{Done}\ (j{:}\alpha)$

Remark: The effect can be reduced as above because $G_K\ G_i \neg \mathit{Intend}_i\ j{:}\alpha \leftrightarrow$ $G \neg \mathit{Intend}_i\ j{:}\alpha$ is valid.

6.3.2.6 Failure ($\langle i, J, K, \mathsf{Failure}, i{:}\alpha\rangle$)

The agent attempted to perform an action but failed. We consider that this act has an Inform part in which he informs that the action has not been done and the agent does not have the intention to do it anymore. But we cannot reduce it to a simple Inform as FIPA-ACL does because it would not be relevant to inform of the failure of an action (and on the resignation of the intention of doing it) if the speaker had not expressed before that he has the intention to

perform the action and if the hearer does not intend that the speaker performs the action. Thus we have to enforce the preconditions of the Inform act:

$\langle i, J, K, \mathsf{Failure}, i{:}\alpha \rangle$

FP: $\neg G_K\ G_i\ (\neg Done\ (i{:}\alpha) \wedge \neg Intend_i\ Done\ (i{:}\alpha))$
$\wedge \neg G_K\ G_J\ (\neg Done\ (i{:}\alpha) \wedge \neg Intend_i\ Done\ (i{:}\alpha))$
$\wedge \neg G_K \neg G_J\ (\neg Done\ (i{:}\alpha) \wedge \neg Intend_i\ Done\ (i{:}\alpha))$
$\wedge \neg G_K \neg G_i\ (\neg Done\ (i{:}\alpha) \wedge \neg Intend_i\ Done\ (i{:}\alpha))$
$G_K\ Intend_j\ Done\ (i{:}\alpha) \wedge G_K\ Intend_i\ Done\ (i{:}\alpha)$

IE: $G_K\ G_i\ (\neg Done\ (i{:}\alpha) \wedge \neg Intend_i\ Done\ (i{:}\alpha))$
$\wedge G_K\ Intend_i\ G_J\ (\neg Done\ (i{:}\alpha) \wedge \neg Intend_i\ Done\ (i{:}\alpha))$
$\wedge G_K\ G_i \neg G_J\ (\neg Done\ (i{:}\alpha) \wedge \neg Intend_i\ Done\ (i{:}\alpha))$
$\wedge G_K \neg Done\ (i{:}\alpha)$

6.3.2.7 Inform Done ($\langle i, J, K, \mathsf{InformDone}, Done_{i:\alpha} \rangle$)

The agent i has performed the action successfully and so informs that he did it. We have to remark that this speech act is useless in our formalism because we consider that actions are public. There is thus no need for agents to inform each other when they performed an action, because other ones are already aware.

We can nevertheless remark that this hypothesis of public action is very strong. It is well adapted for our purpose where we use only dialogical actions. But if we consider also physical actions that can be privately performed (*i.e.* actions for which Axiom $\mathrm{PA}_{I,\alpha}$ does not hold), the informDone speech act should be defined. Following the same reasoning as for Failure, we cannot reduce this act as FIPA-ACL does to Inform that the action was performed and must strengthen the preconditions.

$\langle i, J, K, \mathsf{InformDone}, Done\ (i{:}\alpha) \rangle$

FP: $\neg G_K\ G_i\ Done\ (i{:}\alpha) \wedge \neg G_K\ G_J\ Done\ (i{:}\alpha)$
$\wedge \neg G_K \neg G_J\ Done\ (i{:}\alpha) \wedge \neg G_K \neg G_i\ Done\ (i{:}\alpha)$
$\wedge G_K\ Intend_j\ Done\ (i{:}\alpha) \wedge G\ Intend_i\ Done\ (i{:}\alpha)$

RE: $G_K\ G_i\ Done\ (i{:}\alpha) \wedge G_K\ Intend_i\ G_J\ Done\ (i{:}\alpha) \wedge G_K\ G_i \neg G_J\ Done\ (i{:}\alpha)$

6.3.2.8 Not-Understood

FIPA introduced a special act in the case where there is misperception of the act received by an agent. But we reason on what is publicly grounded and we consider that there is perfect understanding of what is sent. So we do not consider this act.

After having defined our speech acts semantics, we have to prove that our system behaves like the CNP.

6.3.3 Theoretical Results

We here show that our system is complete (*i.e.* we can do all sequences of speech acts allowed by the CNP) and sound (*i.e.* we cannot do more than the allowed sequences) with the CNP. These results entail the termination of the CNP protocol.

6.3.3.1 Soundness and completeness

We suppose that we have a set *FRAME* of frame axioms of the form: $\neg G_K\ Intend_i\ j{:}\beta \rightarrow After_{\langle i,\mathsf{Cfp},j,\alpha\rangle}\ \neg G\ Intend_i\ j{:}\beta$[6] and a set *INIT* containing the relevant preconditions of the form $\neg G_K\ \varphi$ for all the acts appearing in the CNP.

In the sequel we denote by i the initiator of the dialog and by J the set of all participants to the protocol (j represents each participant). To prove soundness and completeness, we begin by showing that the first step of the protocol is executable:

LEMMA. $LAWS \cup FRAME \vdash INIT \rightarrow Happens_{\langle i,J,K,\mathsf{Cfp},\alpha\rangle}\ \top$

PROOF.

By definition of the set INIT, we have the inclusion $FP(\langle i, J, K, \mathsf{Cfp}, \alpha\rangle) \subset$ *INIT and thus the lemma holds.*

□

We can next prove the same lemma for the second step of the protocol.

LEMMA. *For each* $j \in J$, $LAWS \cup FRAME \vdash INIT \rightarrow After_{\langle i,J,K,\mathsf{Cfp},\alpha\rangle}$
$G_K\ (Happens\ (\langle j, i, K, \mathsf{Refuse}, \langle j, i, K, \mathsf{Propose}, j{:}\alpha\rangle\rangle)$
$\vee\ Happens\ (\langle j, i, K, \mathsf{Propose}, j{:}\alpha\rangle))$

PROOF. *Thanks to the effect laws of the call for proposals we have:*
$LAWS \vdash After_{\langle i,J,K,\mathsf{Cfp},\alpha\rangle}\ G_K\ Intend_i\ \bigvee_{j\in J} Done_{\langle j,i,K,\mathsf{Propose},j:\alpha\rangle}\ \top$

Thanks to the frame axioms and the set INITwe have for each $j \in J$*:*
$FRAME \vdash INIT \rightarrow After_{\langle i,J,K,\mathsf{Cfp},\alpha\rangle}$
$\neg G_K\ Intend_i\ \langle j, i, K, \mathsf{Propose}, j{:}\alpha\rangle \wedge \neg G_K\ Intend_j\ \langle j, i, K, \mathsf{Propose}, j{:}\alpha\rangle$

Thus we can derive that:
$LAWS \cup FRAME \vdash INIT \rightarrow After_{\langle i,J,K,\mathsf{Cfp},\alpha\rangle}\ (FP(\langle j, i, K, \mathsf{Refuse}, j{:}\alpha\rangle)$
$\wedge\ FP(\langle j, i, K, \mathsf{Propose}, j{:}\alpha\rangle))$

From that the lemma follows. □

We can prove this kind of lemma for every speech acts sequence of the Contract Net Protocol. They come straightforward from the above characterization of speech acts. Putting all those lemmas together, we can state the following theorem:

THEOREM. *Our system (with our logic and our speech act semantics) is complete with respect to the Contract Net Protocol of FIPA (with the speech act semantic*

[6] Ideally, these axioms should be eliminated by resorting to a solution of the frame problem.

of FIPA), i.e. *every sequence of speech acts $\alpha_1, ..., \alpha_n$ permitted by the protocol is allowed in our system, in the sense that:*
$LAWS \cup FRAME \vdash INIT \rightarrow Happens_{\alpha_1}\ G_K \ldots G_K\ Happens_{\alpha_n}\ \top$

Moreover we prove the converse theorem (which is much more powerful):

THEOREM. *Our system is sound with respect to the Contract Net Protocol of FIPA,* i.e. *no other sequence of speech acts than the ones allowed by the CNP is permitted by our system,* i.e. *if $\alpha_1, .., \alpha_n$ is not admitted by the CNP then:*
$LAWS \cup FRAME \vdash INIT \rightarrow \neg Happens_{\alpha_1}\ G \ldots G\ Happens_{\alpha_n}\ \top$.

PROOF. *We do not present the whole proof here but only highlight its main points.*

Firstly when a kind of speech act is made with a propositional content, it cannot be performed (with the same propositional content) anymore by an agent. This property holds for the two main speech acts, that are Request *and* Inform*: if an agent informs that φ then $G_K\ G_i\ \varphi$ holds, which is inconsistent with the FP of $\langle i, j, K, \mathsf{Inform}, \varphi\rangle$ and of $\langle j, i, K, \mathsf{Inform}, \varphi\rangle$ (and same for* Request*). We can remark that the semantics of the other speech acts is built upon the semantics of the one of* Request *or* Inform*, thus the property holds for all the speech acts.*

Secondly we can prove that the preconditions of the speech acts constrain the possible next acts to the ones permitted by the protocol. □

6.3.3.2 Termination

Our system is sound and complete with the CNP. With its tree-structure, this protocol clearly terminates. So we do not need more evidence to prove that:

THEOREM. *Our system terminates.*

With this result, we avoid infinite cycles that can arise in this type of system.

Given our new semantics of speech acts, we have shown that we can capture the Contract Net Protocol. We have highlighted that the SABS avoids to impose strong hypotheses on agents, such as sincerity, credulity or cooperation. But as these hypotheses can be made in some actual MASs, we will study in the next section how we can integrate them in our logical framework. This will require the addition of some new axioms to our logic.

6.4 How to take into account stronger hypotheses

6.4.1 Sincerity and cooperation

In our formalism, $G_I\ \varphi \rightarrow G_i\ \varphi$ is not valid. Thus, when it is grounded that a piece of information φ holds for agent i (for example after asserting φ) then this does not mean that i indeed believes that φ. So an agent can assert something and believe the contrary. The power of our operator is to a large extent based on

this property. It is possible (but a bit opposed to the main idea of this account) to characterize sincere agents by adding some axioms such as:

$$G_K\, G_i\, \varphi \rightarrow G_i\, \varphi \qquad (Ax_{Sinc})$$

Note that the other way round, $G_i\, \varphi \rightarrow G_K\, G_i\, \varphi$, is not valid either: an agent might privately believe φ while it is not grounded that φ holds for i. Moreover to impose this formula as an axiom would be much more powerful: as soon the agent i believes something, it becomes public that he believes it. Nevertheless this is close to a kind of cooperativity feature of the agent: when an agent believes something, he wants to inform other agents of this fact (Demolombe, 2004), that can be represented by the axiom of strong cooperativity:

$$G_i\, \varphi \wedge G_i\, \neg \mathit{BelIf}_j\, \varphi \rightarrow \mathit{Intend}_i\, \mathit{Done}_{\langle i,j,K,\mathsf{Inform},\varphi\rangle}\, \top \qquad (Ax_{SCoop})$$

Moreover we can consider that agents can be only *publicly sincere*, meaning that they cannot lie about pieces of information that are grounded publicly for them. This can be expressed by the axiom :

$$G_K\, \mathit{Intend}_i\, \mathit{Bel}_j\, \phi(i) \rightarrow G_K\, \phi(i) \qquad (Ax_{publicsincerity})$$

with $\phi(i)$ a mental attitude of the agent i.

6.4.2 Credulity and credibility

Inspiring by (Demolombe, 2004), we can formalize other hypotheses such as the credibility or the credulity. These hypotheses link private mental attitudes of one agent to what others have expressed. In particular we consider that an agent is credulous when he believes everything that has been expressed:

$$G_K\, G_j\, \varphi \rightarrow G_i\, \varphi \qquad (Ax_{Credul})$$

Credulous agents would run into inconsistency when they are agents j_1 and j_2 that have expressed opposite views: $G_K\, G_{j_1}\, \varphi G_K\, G_{j_2}\, \neg\varphi \rightarrow G_i\, \bot$ follows from (Ax_{Credul}). And as we have excluded such inconsistency (by axiom (D_{G_I})), it turns out that a single credulous agents implies unanimity.

Contrarily, an agent is said credible when other agents believe what he has asserted:

$$G_K\, G_i\, \varphi \rightarrow \bigwedge_{k\in K} G_k\, \varphi \qquad (Ax_{Credibl})$$

Note that this axiom is very strong in particular because, thanks to the necessitation rule (RN_{G_I}) and axiom (CG), it allows that a credible agent can ground for the whole group what he asserts:

$$G_K\, G_i\, \varphi \rightarrow G_K\, \varphi$$

6.4.3 Public trust

In many cases, we can safely assume that group J immediately starts to publicly believe an information asserted by agent i, namely when this group trusts the uttering agent in regard to this information.[7] An important particular case is when the group J have asked i to publicly declare that φ (for $J \subseteq K$). This is expressed by the following axiom,:

$$(G_K\ Done_{\langle i,J,K,\mathsf{InformIf},\varphi\rangle}\top \bigwedge_{j\in J} Intend_j\ Done_{\langle i,J,K,\mathsf{InformIf},\varphi\rangle}\top) \rightarrow G_K\ G_J\ \varphi$$

$$(Ax_{HearingIsBelieving})$$

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 outside?"), it becomes grounded that she believes the answer. The intention $Intend_j$ can be triggered with FIPA's *QueryIf* act.

6.5 Conclusion

In this section we have showed how to apply the new ACL semantics approach presented in the previous chapter, that we have called SABS. We have concentrated on the FIPA paradigm, but we could have done exactly the same work for KQML semantics (and the result would have been very close). We began by redefining the semantics of the four FIPA primitives and illustrating it on a case study. Afterwards we apply it to formalize the Contract Net Protocol.

Although our approach is based on the rejection of strong hypotheses required for classical mentalist semantics, we have ended this chapter by introducing additional axioms, allowing the extension of our logic to take them into account. This allows MASs using such constraints to use our framework, but they lose a bit the spirit of the account.

Following our goal of bridging the gap between the two main approaches of ACLs, we present in the following chapter how our logic can also capture the social semantics. We will formalize and discuss Walton & Krabbe's and Colombetti's approaches based on social commitments.

[7] A notorious exception are exam situations.

Chapter 7

Application: formalization of the social approach

7.1 Introduction

After having redefined the semantics of the FIPA mentalist approach of ACLs in the previous chapter, we bridge here the gap between both approaches by redefining social approaches in our logical formalism. We concentrate our work on Walton & Krabbe's and Colombetti's approaches, that are based on the two different kinds of social commitment: propositional commitments (Section 7.2) and commitments in action (Section 7.3).

Our aim is to show that our logical framework allows to formalize both approaches to ACLs. Moreover, as announced in Chapter 5, we aim also at giving a logical characterization of the notion of commitment. It is often used as a primitive to represent social link between agents without having been studied in depth.

7.2 Walton & Krabbe's account

7.2.1 Presentation of W&K theory

As presented in Section 5.4.2.2, Walton & Krabbe (W&K for short) use propositional commitments to manage and describe dialogues. They argue that the semantics of speech acts and the rules of these dialog games describing allowed speech acts sequences can be described in terms of propositional commitments. (Walton and Krabbe, 1995) presents a dialogue type hierarchy based on the notion of conflict. Among them, persuasion dialogues are analyzed in detail, with quite precise descriptions of game rules and speech act semantics in terms of commitments. We now apply our formalism to this particular kind of dialogue.

A persuasion dialogue takes place when there is a conflict between two agents' beliefs.The goal of the persuasion 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 define two kinds of persuasion dialogue: the Permissive Persuasion Dialogue (PPD_0 for short) and the Rigorous Persuasion Dialogue (RPD). RPD is asymmetric (participants have different roles *viz.* proponent and opponent), and is analytic (the initial proposition is decomposed during the dialogue), while PPD_0 is symmetric and non analytic (allows to introduce new arguments).

We show, by characterizing PPD_0, how our formally well-grounded operator can be used to define speech act semantics and game rules instead of the informal commitments *à la* W&K.

In order to simplify our exposition we suppose with W&K that there are only two agents (but the account can easily be generalized to n agents).

7.2.2 Strong and Weak commitments

W&K distinguish two kinds of commitment[1]: those which must be defended by a proof or a justification when challenged, called *assertions*, and those which need not, called *concessions*. We formalize this distinction with the notions of strong commitment ($SC_{i,K}$) and weak commitment ($WC_{i,K}$) of an agent i w.r.t. a group K. Unlike W&K we introduce the set of agents in front of which this commitment has been incurred, *i.e.* agents taking part in the conversation. Indeed this set remains implicit for W&K. It is equivalent to say that the commitment stores are specific to one dialog between members of a group K. These two commitments are linked by the fact that a strong commitment to a proposition implies a weak commitment to it (Walton and Krabbe, 1995, p. 133). In relation with our logical framework, we define:

$$SC_{i,K}\varphi \stackrel{def}{=} G_K\, G_i\, \varphi \qquad (\mathrm{Def}_{SC_{i,K}})$$

$$WC_{i,K}\varphi \stackrel{def}{=} G_K\, \neg G_i\, \neg\varphi \qquad (\mathrm{Def}_{WC_{i,K}})$$

Note that this is an approximation of W&K's *assertion*, noted *a*. Indeed, our $G_K\, G_i\, \varphi$ is "more logical" than W&K's *a(φ)*: W&K allow both $a(\varphi)$ and $a(\neg\varphi)$ to be the case simultaneously, while for us $G_K\, G_i\, \varphi \wedge G_K\, G_i\, \neg\varphi$ is inconsistent. In the case of weak commitment, we agree with W&K's works: $WC_{i,K}\varphi \wedge WC_{i,K}\neg\varphi$ is consistent.

In terms of the preceding abbreviations we can prove:

$$SC_{i,K}\varphi \rightarrow \neg SC_{i,K}\neg\varphi \qquad (\mathrm{D}_{SC_{i,K}})$$

$$SC_{i,K}\varphi \leftrightarrow SC_{i,K}SC_{i,K}\varphi \qquad (4_{SC_{i,K}})$$

$$\neg SC_{i,K}\varphi \leftrightarrow SC_{i,K}\neg SC_{i,K}\varphi \qquad (5_{SC_{i,K}})$$

[1] We can note that authors introduce also a third kind of commitment, that they call *dark-side commitments*. These commitments remain private (and can be unknown by the agent himself) during the dialog until they are revealed. But due to its private feature, we will not consider this kind of commitment.

(D_{G_I}) shows the rationality of the agents: they cannot strongly commit both on φ and $\neg\varphi$. (4_{G_I}) and (5_{G_I}) account for the public character of commitments. With these three theorems, we can show that $SC_{i,K}$ is an operator of a normal modal logic of type KD45, just as G_K.[2]

We can prove additional theorems linking operator G_K with commitments.

$$SC_{i,K}\varphi \leftrightarrow G_K\, SC_{i,K}\varphi \tag{7.1}$$
$$\neg SC_{i,K}\varphi \leftrightarrow G_K\, \neg SC_{i,K}\varphi \tag{7.2}$$
$$WC_{i,K}\varphi \leftrightarrow G_K\, WC_{i,K}\varphi \tag{7.3}$$
$$\neg WC_{i,K}\varphi \leftrightarrow G_K\, \neg WC_{i,K}\varphi \tag{7.4}$$

These four theorems express the public character of the strong and weak commitments. If a commitment holds (resp. does not hold), this is grounded for the group of overhearers. They can be proved from (SR+), (SR−) and definitions of strong and weak commitments.

$$G_K\,\varphi \leftrightarrow SC_{i,K}\, G_K\,\varphi \tag{7.5}$$
$$\neg G_K\,\varphi \leftrightarrow SC_{i,K} \neg G_K\,\varphi \tag{7.6}$$
$$SC_{i,K}\varphi \leftrightarrow SC_{j,K} SC_{i,K}\varphi \tag{7.7}$$
$$\neg SC_{i,K}\varphi \leftrightarrow SC_{j,K} \neg SC_{i,K}\varphi \tag{7.8}$$

(7.5) and (7.6) entail that it is equivalent to say that a formula is grounded (resp. ungrounded) and that the agents are committed to this grounded (resp. ungrounded) formula. (7.7) and (7.8) mean that each agent is committed to the other agents' commitments, and non-commitments.

The following theorems links strong and weak commitments.

$$SC_{i,K}\varphi \rightarrow WC_{i,K}\varphi \tag{7.9}$$
$$WC_{i,K}\varphi \rightarrow \neg SC_{i,K}\neg\varphi \tag{7.10}$$

(7.9) says that strong commitment implies weak commitment. (7.10) expresses that if agent i is weakly committed to φ then i is not strongly committed to $\neg\varphi$.

$$WC_{i,K}\varphi \leftrightarrow SC_{j,K}\, WC_{i,K}\varphi \tag{7.11}$$
$$\neg WC_{i,K}\varphi \leftrightarrow SC_{j,K} \neg WC_{i,K}\varphi \tag{7.12}$$

(7.11) expresses that weak commitment is public. (7.12) is similar for absence of weak commitment.

[2] From the definition of G_K, we can prove that the axiom K is a theorem for $SC_{i,K}$ and that the rule of necessitation RN can be applied to it.

7.2.3 Formalization of PPD_0

For speech acts we keep the notation introduced in the previous chapter: speech acts are 5-tuples of the form $\langle i, J, K, FORCE, \varphi \rangle$ where $i \in AGT$ is the *author* of the speech act (*i.e.* the speaker), $K \subseteq AGT$ the *group* of attendees, $J \subseteq K \setminus \{i\}$ the set of its *addressees*, $FORCE$ its illocutionary force, and φ a formula denoting its propositional content. As said above we consider dialogues between only two agents, thus the addressees group J is the singleton $\{j\}$, and the attendees group K is $\{i, j\}$. In the sequel we will use a simplified notation for speech acts: $\langle s, h, FORCE, p \rangle$, where s is the speaker and h the hearer. Moreover to stay close to W&K's formalism, p denotes the propositional content.

The dialogues that we want to formalize (W&K-like dialogues) are controlled by some conventions: the rules of the game, which describe the allowed sequences of speech acts. The allowed sequences of acts are those of W&K's PPD_0 (Walton and Krabbe, 1995, p. 150-151). They are formalized in Figure 7.1 and will be discussed below. For example, after a speech act $\langle s, \mathsf{Assert}, h, p \rangle$,

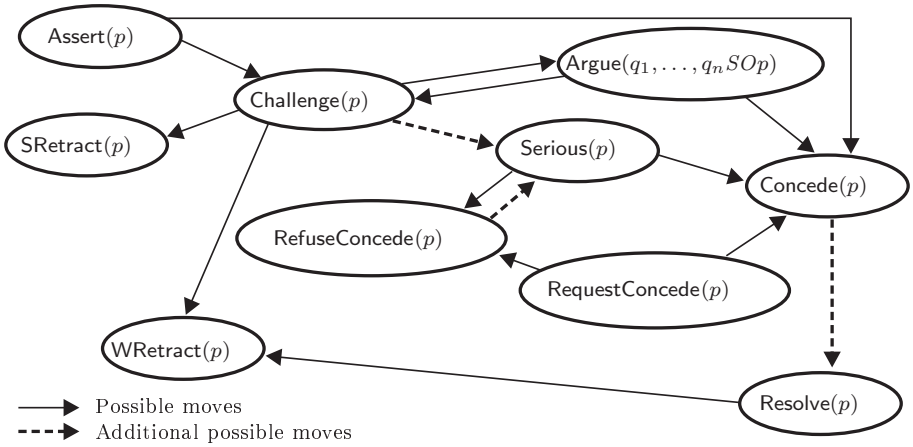

Figure 7.1: (Additional) possible moves after each act

the hearer can only challenge p or concede p. We formalize W&K speech acts 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.* Figure 7.1) and the other one is the set of acts the hearer can perform in response (*cf.* Figure 7.2).

We suppose that initially nothing is grounded, *i.e.* we reuse the set *INIT* introduced Section 6.3.3 of the above chapter, adapted for PPD_0 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.

Precond(α)	Act α	Effects(α)
$\neg SC_s\, p$	$\langle s, \mathsf{Assert}, h, p\rangle$	$SC_s\, p$
$SC_s\, p$	$\langle s, h, \mathsf{SRetract}, p\rangle$	$\neg SC_s\, p$
$WC_s\, p$	$\langle s, h, \mathsf{WRetract}, p\rangle$	$\neg\, WC_s\, p$
$SC_s\, p \wedge \neg\, WC_h\, p$	$\langle s, h, \mathsf{Argue}, q_1, ..., q_n SOp\rangle$	$\bigwedge_{1\leq i\leq n} SC_s\, q_i \wedge$ $SC_s\, (\bigwedge_{1<i\leq n} q_i \rightarrow p)$
$\neg\, WC_s\, p$	$\langle s, h, \mathsf{Concede}, p\rangle$	$WC_s\, p$
$\neg\, WC_s\, p$	$\langle s, h, \mathsf{refuseConcede}, q\rangle$	$\neg\, WC_s\, p$
$SC_s\, q \wedge \neg\, WC_h\, q \wedge$ $\neg\, WC_h\, p$	$\langle s, h, \mathsf{requestConcede}, p\rangle$	$\emptyset$
$\neg\, WC_s\, p \wedge SC_h\, p \wedge$ $\neg GDone_{\langle s,h,\mathsf{Challenge},p\rangle}\, \top$	$\langle s, h, \mathsf{Challenge}, p\rangle$	$\emptyset$
$\neg\, WC_h\, p$	$\langle s, h, \mathsf{Serious}, p\rangle$	$\emptyset$
$WC_h\, p \wedge WC_h\, q \wedge$ $(p \leftrightarrow \neg q)$	$\langle s, h, \mathsf{Resolve}, p\rangle$	$\emptyset$

Table 7.1: Preconditions and effects of speech acts (with commitments).

To Concede p means to admit that p could hold, where p is a strong commitment of the other party (e.g. p has been asserted). 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 seems weaker than 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 content this act can only be performed once.

To defend a challenged assertion p, an argument, expressed by Argue, must have p as conclusion and a set of propositions $q_1 ... q_n$ as premises. We write it as follows:

$$q_1 ... q_n SOp \stackrel{def}{=} q_1 \wedge ... \wedge q_n \wedge (q_1 \wedge ... \wedge q_n \rightarrow p) \qquad (\mathrm{Def}_{SO})$$

The effect of this act is that the speaker is strongly committed on all premises q_1, ..., q_n and on the implication $q_1 \wedge ... \wedge q_n \rightarrow p$. 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 digressions W&K suppose that an unchallenged assertion cannot be defended by an argument. Note that, we took over their form of the support of arguments, *viz.* $\varphi \rightarrow \psi$, although we are aware that more complex forms of reasoning occur in real world argumentation.

Acts α	Constraints on the possible actions following α
$\langle s, \mathsf{Assert}, h, p\rangle$	$G_K\,(Choice_h\ Happens\ (\langle h, s, \mathsf{Challenge}, p\rangle) \vee$ $Choice_h\ Happens\ (\langle h, s, \mathsf{Concede}, p\rangle))$
$\langle s, h, \mathsf{SRetract}, p\rangle$	$\emptyset$
$\langle s, h, \mathsf{WRetract}, p\rangle$	$\emptyset$
$\langle s, h, \mathsf{requestConcede}, p\rangle$	$G_K\,(Choice_h\ Happens\ (\langle h, s, \mathsf{refuseConcede}, p\rangle) \vee$ $Choice_h\ Happens\ (\langle h, s, \mathsf{Concede}, p\rangle))$
$\langle s, h, \mathsf{Argue}, q_1, ..., q_n SOp\rangle$	$\bigwedge_{1\leq i\leq n} G_K\,(Choice_h\ Happens\ (\langle h, s, \mathsf{Challenge}, q_i\rangle) \vee$ $Choice_h\ Happens\ (\langle h, s, \mathsf{Concede}, q_i\rangle))$ $\wedge$ $G_K\,(Choice_h\ Happens\ (\langle h, s, \mathsf{Challenge}, q_1 \wedge ... \wedge q_n \rightarrow p\rangle) \vee$ $Happens\ (\langle h, s, \mathsf{Concede}, q_1 \wedge ... \wedge q_n \rightarrow p\rangle))$
$\langle s, h, \mathsf{Challenge}, p\rangle$	$G_K\,(Choice_h\ Happens\ (\langle h, s, \mathsf{SRetract}, p\rangle) \vee$ $Choice_h\ Happens\ (\langle h, s, \mathsf{WRetract}, p\rangle) \vee$ $Choice_h\ Happens\ (\langle h, s, \mathsf{Argue}, q_1, ..., q_n SOp\rangle) \vee$ $Choice_h\ Happens\ (\langle h, s, \mathsf{Serious}, p\rangle))$
$\langle s, h, \mathsf{Concede}, p\rangle$	$\emptyset$
$\langle s, h, \mathsf{refuseConcede}, p\rangle$	$\emptyset$
$\langle s, h, \mathsf{Serious}, p\rangle$	$G_K\,(Choice_h\ Happens\ (\langle h, s, \mathsf{refuseConcede}, p\rangle) \vee$ $Choice_h\ Happens\ (\langle h, s, \mathsf{Concede}, p\rangle))$
$\langle s, h, \mathsf{Resolve}, p\rangle$	$G_K\,(Choice_h\ Happens\ (\langle h, s, \mathsf{WRetract}, p\rangle) \vee$ $Choice_h\ Happens\ (\langle h, s, \mathsf{WRetract}, \neg p\rangle))$

Table 7.2: Additional effects of speech acts.

At any time, the speaker may request more concessions (with a RequestConcede 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 we argue that it is not indeed the same kind of act, and we choose to introduce two different acts: $\langle s, h, \mathsf{WRetract}, p\rangle$ to retract his own weak commitment on p, and $\langle s, h, \mathsf{refuseConcede}, p\rangle$ to decide not to concede p. 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, $WC_{i,K}\varphi \wedge WC_{i,K}\neg\varphi$ is satisfiable, but not $SC_{i,K}\varphi \wedge SC_{i,K}\neg\varphi$. 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[3]. When a party detects such a contradiction in the other party's commitment store, he can ask him to resolve it (with the act Resolve(p,q) where "p and q are explicit contradictories" (Walton and Krabbe, 1995, p. 151)). The other party

[3]W&K allow the agents to have some contradictory concessions (WC) and assertions (SC) in their commitment store (*i.e.* $SC_{i,K}\varphi$ and $SC_{i,K}\neg\varphi$ or $WC_{i,K}\varphi$ and $WC_{i,K}\neg\varphi$ can hold simultaneously).

must retract one of the inconsistent propositions.

W&K do not make any inference in the commitment 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 opponent. In our formalism, the act Resolve applies only to weak commitments. Moreover the two contradictory weak commitments cannot be derived from two inconsistent strong commitments (which W&K allow), because strong commitments are consistent in our logic (theorem $\mathrm{D}_{SC_{i,K}}$).

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.

As we have already said, W&K define a third commitment store that contains what they call *dark-side commitments*. Whereas assertions and concessions (light-side commitments) are public, no agent is necessarily aware of these commitments. They characterize deep features of agents. 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[4].

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.

7.2.4 Example

We recast an example of a persuasion dialogue given by W&K (Walton and Krabbe, 1995, p. 153) to illustrate the dialogue game PPD_0 (see Figure 7.2): initially, agent i asserts p_1 and agent j asserts p_2. Thus, the following preparatory moves have been performed: $\langle i, \mathsf{Assert}, j, p_1\rangle$ and $\langle j, \mathsf{Assert}, i, p_2\rangle$. After each move, the agents' commitment stores are updated (see Table 7.3). In his first move, j asks i to concede p_3 and challenges p_1. i responds by conceding p_3, and so on. In move 7., agent j concedes p_1 which is the thesis of his opponent. 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.

As we have said above, in order to stay consistent with our logical framework,

[4]Thus W&K consider agents can be publicly inconsistent but they cannot hide deep dark commitments if the opponent insists with a Serious. We note in passing that this feature of W&K's approach is subject to the same criticism as the mentalist approaches as exposed in Section 5.3.5 of Chapter 5: just as private beliefs, dark-side commitments cannot be verified, and a participant in a conversation can never exclude that the dark-side commitments of another participant is empty.

1. $\langle j, i, \mathsf{requestConcede}, p_3\rangle$,
   $\langle j, i, \mathsf{Challenge}, p_1\rangle$
2. $\langle i, j, \mathsf{Concede}, p_3\rangle$,
   $\langle i, j, \mathsf{Serious}, p_1\rangle$,
   $\langle i, j, \mathsf{Argue}, p_3 SO p_1\rangle$,
   $\langle i, j, \mathsf{Challenge}, p_2\rangle$
3. $\langle j, i, \mathsf{refuseConcede}, p_1\rangle$,
   $\langle j, i, \mathsf{Concede}, p_3 \rightarrow p_1\rangle$,
   $\langle j, i, \mathsf{Argue}, p_4, p_5 SO p_2\rangle$,
   $\langle j, i, \mathsf{Challenge}, p_3\rangle$
4. $\langle i, j, \mathsf{Concede}, p_5\rangle$,
   $\langle i, j, \mathsf{Concede}, p_4 \wedge p_5 \rightarrow p_2\rangle$,
   $\langle i, j, \mathsf{Serious}, p_3\rangle$,
   $\langle i, j, \mathsf{Argue}, \neg p_4, p_5 SO p_3\rangle$,
   $\langle i, j, \mathsf{Challenge}, p_4\rangle$
5. $\langle j, i, \mathsf{WRetract}, p_3 \rightarrow p_1\rangle$,
   $\langle j, i, \mathsf{Concede}, p_3\rangle$,
   $\langle j, i, \mathsf{Concede}, \neg p_4\rangle$,
   $\langle j, i, \mathsf{Concede}, \neg p_4 \wedge p_5 \rightarrow p_3\rangle$,
   $\langle j, i, \mathsf{Argue}, p_3 SO p_4\rangle$,
   $\langle j, i, \mathsf{Challenge}, p_3 \rightarrow p_1\rangle$
6. $\langle i, j, \mathsf{Resolve}, p_4\rangle$,
   $\langle i, j, \mathsf{Argue}, \neg p_4 SO p_3 \rightarrow p_1\rangle$,
   $\langle i, j, \mathsf{Challenge}, p_3 \rightarrow p_4\rangle$
7. $\langle j, i, \mathsf{WRetract}, p_4\rangle$,
   $\langle j, i, \mathsf{WRetract}, p_3 \rightarrow p_4\rangle$,
   $\langle j, i, \mathsf{SRetract}, p_5\rangle$,
   $\langle j, i, \mathsf{SRetract}, p_3\rangle$,
   $\langle j, i, \mathsf{WRetract}, p_4 \wedge p_5 \rightarrow p_2\rangle$,
   $\langle j, i, \mathsf{Concede}, \neg p_4 \rightarrow (p_3 \rightarrow p_1)\rangle$,
   $\langle j, i, \mathsf{Concede}, p_3 \rightarrow p_1\rangle$,
   $\langle j, i, \mathsf{Concede}, p_1\rangle$,
   $\langle j, i, \mathsf{Argue}, p_6 SO p_2\rangle$,

Figure 7.2: Example of dialogue (from (Walton and Krabbe, 1995, p. 153))

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 abandon the paraconsistent aspects of W&K, viz. that an agent can have assertions or concessions that are jointly inconsistent, in order to keep in line with standard properties of the modal operator G_K .

Now we can establish formally that our logic captures W&K's PPD_0-dialogues. For example we have:

THEOREM.

$$LAWS \models INIT \rightarrow After_{\langle s,\mathsf{Assert},h,p\rangle}\,((\neg WC_h\, p \wedge \neg Done_{\langle h,s,\mathsf{Challenge},p\rangle}\,\top) \rightarrow G_K\,(Happens\,(\langle h, s, \mathsf{Challenge}, p\rangle) \vee Happens\,(\langle h, s, \mathsf{Concede}, p\rangle)))$$

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 7.2) the formula:*

$$After_{\langle s,\mathsf{Assert},h,p\rangle}\; G_K\,(Choice_h\; Happens\,(\langle h, s, \mathsf{Challenge}, p\rangle) \vee Choice_h\; Happens\,(\langle h, s, \mathsf{Concede}, p\rangle))$$

The precondition for $\langle h, s, \mathsf{Challenge}, p\rangle$ *is:*

$$\neg WC_h\, p \wedge SC_s\, p \wedge \neg Done_{\langle h,s,\mathsf{Challenge},p\rangle}\,\top$$

Grounded propositions	SC_i	WC_i	SC_j	WC_j
$\emptyset$	p_1		p_2	
$WC_{i,K}p_3$ $SC_{i,K}p_3$, $SC_{i,K}p_3 \rightarrow p_1$	p_1, p_3, $p_3 \rightarrow p_1$			
$WC_{j,K}p_3 \rightarrow p_1$, $SC_{j,K}p_4$, $SC_{j,K}p_5$, $SC_{j,K}p_4 \wedge p_5 \rightarrow p_2$			p_2, p_4, p_5, $p_4 \wedge p_5 \rightarrow p_2$	$p_3 \rightarrow p_1$
$WC_{i,K}p_5$, $SC_{i,K}\neg p_4$, $SC_{i,K}p_5$, $SC_{i,K}\neg p_4 \wedge p_5 \rightarrow p_3$, $WC_{i,K}p_4 \wedge p_5 \rightarrow p_2$	p_1, $p_3 \rightarrow p_1$, p_3, $\neg p_4$, p_5, $\neg p_4 \wedge p_5 \rightarrow p_3$	$p_4 \wedge p_5 \rightarrow p_2$		
$\neg WC_{j,K}p_3 \rightarrow p_1$, $WC_{j,K}p_3$, $WC_{j,K}\neg p_4 \wedge p_5 \rightarrow p_3$, $SC_{j,K}p_3$, $SC_{j,K}p_3 \rightarrow p_4$, $WC_{j,K}\neg p_4$			p_2, p_4, p_5, p_3, $p_3 \rightarrow p_4$, $p_4 \wedge p_5 \rightarrow p_2$	$\neg p_4$, $\neg p_4 \wedge p_5 \rightarrow p_3$
$SC_{i,K}\neg p_4$, $SC_{i,K}\neg p_4 \rightarrow (p_3 \rightarrow p_1)$	p_3, $p_3 \rightarrow p_1$, p_1, $\neg p_4$, p_5, $\neg p_4 \wedge p_5 \rightarrow p_3$, $\neg p_4 \rightarrow (p_3 \rightarrow p_1)$	$p_4 \wedge p_5 \rightarrow p_2$		
$\neg SC_{j,K}p_4$, $\neg WC_{j,K}p_4$ $\neg WC_{j,K}p_3 \rightarrow p_4$, $\neg SC_{j,K}p_3$, $\neg SC_{j,K}p_5$, $\neg SC_{j,K}p_3 \rightarrow p_4$, $\neg WC_{j,K}p_4 \wedge p_5 \rightarrow p_2$, $\neg SC_{j,K}p_4 \wedge p_5 \rightarrow p_2$ $WC_{j,K}p_3 \rightarrow p_1$, $WC_{j,K}p_1$, $WC_{j,K}\neg p_4 \rightarrow (p_3 \rightarrow p_1)$ $SC_{j,K}p_6$, $SC_{j,K}p_6 \rightarrow p_2$			p_2, p_6, $p_6 \rightarrow p_2$	$\neg p_4$, $\neg p_4 \rightarrow (p_3 \rightarrow p_1)$, p_3, p_5, $p_3 \rightarrow p_1$, p_1, $\neg p_4 \wedge p_5 \rightarrow p_3$

Table 7.3: Commitment stores in the example dialogue

Now the postcondition of $\langle s, \mathsf{Assert}, h, p\rangle$ *is* $SC_s\, p$. *Hence we have by the law of intentional action (*Int_{Choice_i, α_i}*):*

$$LAWS \models After_{\langle s,\mathsf{Assert},h,p\rangle}\,(\neg WC_h\, p \wedge \neg Done_{\langle h,s,\mathsf{Challenge},p\rangle}\, \top \rightarrow (Choice_h\ Happens\ (\langle h, s, \mathsf{Challenge}, p\rangle) \rightarrow Happens\ (\langle h, s, \mathsf{Challenge}, p\rangle)))$$

Similarly, for concede we have:

$$LAWS \models INIT \rightarrow After_{\langle s,\mathsf{Assert},h,p\rangle}\,(\neg WC_h\, p \rightarrow (Choice_h\ Happens\ (\langle h, s, \mathsf{Concede}, p\rangle) \rightarrow Happens\ (\langle h, s, \mathsf{Concede}, p\rangle)))$$

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, then 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_K is a normal modal logic, and thus validates $(SC_{i,K}p \wedge SC_{i,K}q) \rightarrow SC_{i,K}(p \wedge q)$.

We finally consider in the next section Colombetti *et al.*'s account of commitments. It is more oriented toward ACLs semantics definition. Moreover authors argue that their account can take into account both commitments in proposition and in action. Its study will thus permit us to extend the formalization of propositional commitments presented above in more general cases and to introduce a formalization of commitments in action. Thus we will be able to give formal relations linking both kinds of commitments.

7.3 Colombetti *et al.*'s account

As remarked in Section 5.4.3.1, Colombetti *et al.* manage commitments *via* the various states it can hold (described in the commitment life-cycle in Figure 5.1 of Chapter 5). The authors argue that this life-cycle is adapted for all kinds of commitment and in particular also for propositional commitments. We have showed that it should be discussed and adapted to take them into account as we will show in the following section before studying commitments in action.

In the sequel, we stay close to Castelfranchi's view of the links between commitments and actions: we base our formalization mainly on speech act theory by characterizing commitments in terms of speech acts inducing them. We mainly study the primitive assertive, that is Assert. Assertive speech acts induce pending propositional commitments. Likewise directive ones induce unset action commitments, and commissive speech acts induce pending action commitments. The primitives are Direct and Commit.

To simplify the exposition, we here drop overhearers introduced in the previous section, which means that we ignore witnesses. In Colombetti's account commitments are for a part described by their state. Note that when we will use the generic term "social commitment", we will refer to a pending commitment.

7.3.1 Propositional commitment

For description of propositional commitments, we reuse some definitions introduced for W&K's persuasion dialogues. Note that we are not limited to this particular type of dialogue here and thus all these definitions can be used in every kind of dialogue.

We need to simplify Colombetti's general commitment life-cycle and adapt it to propositional commitments. In particular we drop the *unset* state that we do not consider relevant for them, as stated in section 5.4.3.1. We consider that transitions between states are not the result of general institutional actions or events as in (Fornara and Colombetti, 2002), but rather of speech acts of which we will give the semantics.

7.3.1.1 *Pending* commitment

For W&K and Colombetti, propositional commitments come from an assertive speech act. Following speech act theory (Vanderveken, 1990), Assert is the primitive assertive speech act. It has "the preparatory condition that the speaker has reasons or evidence for the truth of the propositional content" and "the sincerity condition that the speaker believes the propositional content" (Vanderveken, 1990, p. 125). Contrarily to Inform speech acts defined in the previous chapter, Assert neither supposes anything about the group of hearers (*i.e.* that they are not aware of φ), nor about a speaker's intention that the hearer learns something. Thus we only impose as precondition that the speaker stays consistent, *i.e.* that he has neither asserted nor conceded the contrary. With the terminology used to represent W&K's account that means that he has neither incurred a strong nor a weak commitment on $\neg\varphi$. The Feasible Precondition is thus reduced to $\neg G_{\{i,j\}}\,\neg\varphi$, because $\neg\, WC_{i,\{i,j\}}\neg\varphi \rightarrow \neg SC_{i,\{i,j\}}\neg\varphi$ (thanks to Theorem 7.9).

As Illocutionary Effect, we impose (with Vanderveken (Vanderveken, 1990)) that the speaker expresses the preparatory and the sincerity condition. We simplify this account by omitting the preparatory condition. We consider that the existence of reasons and evidence for the truth of φ are implicitly embedded in the fact that the speaker believes that φ holds.

We thus define the Assert speech act by:

$$\begin{array}{l}\langle i, \mathsf{Assert}, j, \varphi\rangle \\ \quad \text{FP: } \neg G_{\{i,j\}}\,\neg G_i\,\varphi \\ \quad \text{IE: } G_{\{i,j\}}\,G_i\,\varphi\end{array}$$

Note that our account conforms to the speech act theory as to the links between Assert and Inform speech acts, *viz.* that Inform is an extension of Assert. The Assert precondition (resp. postcondition) is implied by the Inform precondition (resp. postcondition).

As mentioned above, a propositional pending commitment is the result of an assertion. We therefore identify a pending propositional commitment with Strong Commitments defined in the previous chapter. Thus we have the following definition of a pending commitment C_{prop}:

DEFINITION.

$$PC_{i,j}(\varphi) \stackrel{def}{=} C_{prop}(pending, i, j, \varphi) \stackrel{def}{=} G_{\{i,j\}}\,G_i\,\varphi$$

We can remark that due to the equivalence $G_i\,\varphi \leftrightarrow G_i\,G_i\,\varphi$ (Theorem (4_{G_I})), we have the following theorem:

THEOREM. $\models C_{prop}(pending, i, j, \varphi) \leftrightarrow C_{prop}(pending, i, j, G_i\,\varphi)$

This theorem means that, by performing a speech act, an agent expresses the sincerity conditions, that is in this case that he believes what he has asserted.

7.3.1.2 *Canceled* commitment

A canceled commitment is a pending commitment that has been retracted. It is thus the result of a Cancel speech act that can be identified with the act named SRetract of W&K's formalization (Section 7.2.3). This act drops the strong commitment ($G_{\{i,j\}}\,G_i\,\varphi$).

$$\begin{aligned}&\langle i, j, \mathsf{Cancel}, \varphi\rangle\\ &\quad\text{FP: } G_{\{i,j\}}\,G_i\,\varphi\\ &\quad\text{IE: } \neg G_{\{i,j\}}\,G_i\,\varphi\end{aligned}$$

Thus an agent has a canceled commitment when he has canceled (*i.e.* $\neg G_{\{i,j\}}\,G_i\,\varphi$ holds) a pending commitment incurred in the past ($PG_{\{i,j\}}\,G_i\,\varphi$), under the condition that $\neg\varphi$ has not been conceded (this last condition makes the distinction between canceled and violated commitments):

DEFINITION.

$$C_{prop}(canceled, i, j, \varphi) \stackrel{def}{=} PG_{\{i,j\}}\,G_i\,\varphi \wedge \neg G_{\{i,j\}}\,G_i\,\varphi \wedge \neg G_{\{i,j\}}\,\neg G_i\,\varphi$$

7.3.1.3 *Violated* commitment

To characterize a violated commitment, we extend the loss PPD_0-rule: an agent has violated commitments when he has retracted it and, as additional condition, when he has conceded the converse. This represents the case of a public contradiction[5].

DEFINITION.

$$C_{prop}(violated, i, j, \varphi) \stackrel{def}{=} PG_{\{i,j\}}\,G_i\,\varphi \wedge G_{\{i,j\}}\,\neg G_i\,\neg\varphi \wedge G_{\{i,j\}}\,\neg G_i\,\varphi$$

[5]Our logical framework does not allow to have contradictory commitments ($C(pending, i, j, \varphi)$ and $C(pending, i, j, \neg\varphi)$). Hence an agent has to drop his commitment before conceding the contrary. This state can be reached for example by the following course of actions:

$[\langle i, j, \mathsf{SRetract}, \varphi\rangle][\langle i, j, \neg\varphi, \mathsf{Concede},]\rangle\, G_{\{i,j\}}\,\neg G_i\,\varphi$

7.3.1.4 *Fulfilled* commitment

A proponent i wins a PPD_0 game when the opponent j concedes his thesis. Thus a PPD_0 game is won by an agent i iff $G_{\{i,j\}}\, G_i\, \varphi \wedge G_{\{i,j\}}\, \neg G_j\, \neg\varphi$ holds.

DEFINITION.

$$C_{prop}(\mathit{fulfilled}, i, j, \varphi) \stackrel{def}{=} G_{\{i,j\}}\, G_i\, \varphi \wedge G_{\{i,j\}}\, \neg G_j\, \neg\varphi$$

Properties. Our propositional commitment has the following properties:

$$\frac{\vdash \varphi \leftrightarrow \psi}{\vdash PC_{i,j}(\varphi) \leftrightarrow PC_{i,j}(\psi)} \text{ and } \frac{\vdash \varphi \leftrightarrow \psi}{\vdash C_{prop}(\mathit{fulfilled}, i, j, \varphi) \leftrightarrow C_{prop}(\mathit{fulfilled}, i, j, \psi)}$$

But we do **not** have:

$$\frac{\vdash PC_{i,j}(\varphi) \leftrightarrow PC_{i,j}(\psi)}{\vdash C_{prop}(\mathit{fulfilled}, i, j, \varphi) \leftrightarrow C_{prop}(\mathit{fulfilled}, i, j, \psi)}$$

In particular we have $PC_{i,j}(\varphi) \leftrightarrow PC_{i,j}(G_i\, \varphi)$ but not $C_{prop}(\mathit{fulfilled}, i, j, \varphi) \leftrightarrow C_{prop}(\mathit{fulfilled}, i, j, G_i\, \varphi)$, because $\nvdash \varphi \leftrightarrow G_i\, \varphi$.

We have thus made clearer the propositional commitment notion of Colombetti's account and showed that it is indeed a generalization of the W&K's account. This formal characterization will also allow us to link propositional commitments with commitments in action that we will define in the following section.

7.3.2 Commitment in action

For commitments in action, we stay close to Colombetti's life-cycle as it appears to be well adapted for this kind of commitments. We will not consider transitions via low-level institutional actions, but via high-level speech acts.

7.3.2.1 *Pending* commitment

Following Colombetti, there are two ways to incur a pending commitment in action: either by committing oneself spontaneously with *e.g.* a Promise, or by accepting to commit on a requested action. Thus in both cases, it is the result of a commissive speech act. As above we consider the primitive commissive speech act that is Commit. A commissive speech act has "the condition that the propositional content represents a future course of action of the speaker", "the preparatory condition that the speaker is capable of carrying out that action" and "the sincerity condition that he intends to carry it out" (Vanderveken, 1990, p. 125–126). Thus by performing a commissive speech act, an agent expresses at least that he has the intention to perform the action ($G_{\{i,j\}}\, \mathit{Intend}_i\, \mathit{Done}_{i:\alpha}\, \top$)

and that he believes that he can carry out the action ($G_{\{i,j\}}\, \neg G_i\, \neg F Done_{i:\alpha}\, \top$). We can remark that:

LEMMA. $G_{\{i,j\}}\, Intend_i\, Done_{i:\alpha}\, \top \rightarrow G_{\{i,j\}}\, \neg G_i\, \neg F Done_{i:\alpha}\, \top$

PROOF.

1. $\vdash Intend_i\, \varphi \stackrel{def}{=} Choice_i\, FG_i\, \varphi \wedge \neg G_i\, \varphi \wedge \neg G_i\, FG_i\, \varphi$, *by Definition* (Def_{Intend_i})
2. $\vdash Intend_i\, \varphi \rightarrow Choice_i\, FG_i\, \varphi$, *by 1. and LP*
3. $\vdash Choice_i\, \varphi \rightarrow \neg G_i\, \neg \varphi$, *by Axiom* $(BA1_{G_i, Choice_i})$ *and* (D_{Choice_i})
4. $\vdash G_i\, Done_{i:\alpha}\, \top \leftrightarrow Done_{i:\alpha}\, \top$, *by Axioms* $PA_{I,\alpha}$ *and* $NA_{I,\alpha}$
5. $\vdash Intend_i\, Done_{i:\alpha}\, \top \rightarrow \neg G_i\, \neg F Done_{i:\alpha}\, \top$, *by 2., 3., 4. and LP*
6. $\vdash G_{\{i,j\}}\, Intend_i\, Done_{i:\alpha}\, \top \rightarrow G_{\{i,j\}}\, \neg G_i\, \neg F Done_{i:\alpha}\, \top$, *by 5. and* RN_{G_I}

□

As precondition, we only impose that the agent stays coherent (*i.e.* that the performance of this action does not involve inconsistency). Thus we have:

$\langle i, \mathsf{Commit}, j, Done_{i:\alpha}\, \top \rangle$
FP: $\neg G_{\{i,j\}}\, \neg Intend_i\, Done_{i:\alpha}\, \top$
IE: $G_{\{i,j\}}\, Intend_i\, Done_{i:\alpha}\, \top$

We can thus characterize a pending commitment in action.

DEFINITION.

$$C(pending, i, j, \alpha) \stackrel{def}{=} G_{\{i,j\}}\, Intend_i\, Done_{i:\alpha}\, \top$$

We can link this formal definition with the above discussion on the nature of the commitment in action (*cf.* Section 5.4.2.1). We had argued that it should be publicly grounded that the debitor i has the intention to perform the action and that nothing is imposed on the choice of the creditor j. Finally even though we admit the importance of the deontic aspect in social commitments, in this attempt of formalization oriented toward ACL applications, we omit the deontic part of the commitment. Thus the above definition matches with the informal characterization given in the previous chapter.

In that respect, note that Promise is often used to represent commissive speech acts. Being a particular commissive act, it inherits the preconditions and postconditions of Commit. But it brings also into play the creditor's goal and obligations ((Searle, 1969) and (Vanderveken, 1990, p. 182)). Thus its precise formalization remains out of the scope of this chapter.

Links with propositional commitments. As far as we are aware, no link has been identified in the literature up to now between propositional commitment and commitment in action. By uttering : "I promise to take out the garbage", John is incurring a commitment to take out the garbage. Moreover, following Speech Act Theory (Vanderveken, 1990), he is also expressing at least that he has the intention to perform the action to take out the garbage. Thus he is also committed on this proposition. As said in Section 5.4.2.2 he can be sanctioned if he does not perform this action, but also if he utters that he does not have this intention (the sanction would be at the dialogue level: he would appear to be incoherent and thus untrustworthy). As the formalization shows, by incurring commitment on action[6], an agent incurs *de facto* some propositional commitments. In this example, John is committed to his intention to perform the action.

Due to the equivalence: $Intend_i\,\varphi \leftrightarrow G_i\,Intend_i\,\varphi$ (see Section 3.3.5), when an pending commitment in action is incurred, a propositional commitment appears:

THEOREM. $\models C(pending, i, j, \alpha) \leftrightarrow C_{prop}(pending, i, j, Intend_i\,Done_{i:\alpha}\,\top)$

7.3.2.2 *Unset* commitment

An unset commitment of j toward i represents the particular social relation that results from the performance by agent i of a request, an order or another directive speech act. To formalize such a commitment, we need to consider what is primitive in directive speech acts and to formalize the primitive directive speech act: that is Direct. Directives are close to commissives: the distinction is only on the author of the action (Vanderveken, 1990). A directive speech act has "the condition that the propositional content represents a future course of action of the hearer", "the preparatory condition that the hearer can carry out that action" and "the preparatory condition that the speaker desires or wants the hearer to carry it out" ((Vanderveken, 1990, p. 126)). Thus by performing a directive speech act, an agent expresses at least that he has the intention that the hearer performs the action ($G_{\{i,j\}}\,Intend_i\,Done_{j:\alpha}\,\top$), and that he believes that the hearer can carry out the action ($G_{\{i,j\}}\,\neg G_i\,\neg F Done_{j:\alpha}\,\top$). As previously the first formula implies the second one. As precondition, we only impose that the agent stays coherent with previous commitments. Thus we have:

$\langle i, \mathsf{Direct}, j, Done_{j:\alpha}\,\top\rangle$
FP: $\neg G_{\{i,j\}}\,\neg Intend_i\,Done_{j:\alpha}\,\top$
IE: $G_{\{i,j\}}\,Intend_i\,Done_{j:\alpha}\,\top$

We can now characterize an unset commitment:

[6] We consider here that agents only incur commitments intentionally by performing speech acts. We do not take into account commitments coming automatically, *e.g.* by representatives or due to social position... (see (Walton and Krabbe, 1995, p. 32), for a detailed account of the ways of incurring commitments).

DEFINITION.

$$C(unset, j, i, \alpha) \stackrel{def}{=} G_{\{i,j\}}\, Intend_i\, Done_{j:\alpha}\, \top$$

We note that contrarily to Colombetti, it is directed from i toward j. We consider that after a request of i it is doubtful that agent j is a debtor of a kind of commitment. We consider that in this case, only i is committed to something.

Link with propositional commitments. As above, this commitment is equivalent to a propositional commitment:

THEOREM. $C(unset, j, i, \alpha) \leftrightarrow C(pending, i, j, Intend_i\, Done_{j:\alpha}\, \top)$

For example, when Mary requests John to take out the garbage, she expresses she wants John to perform the action *takeOutGarbage*, *i.e.* $G_{\{m,j\}}\, Intend_m\, Done_{j:takeOutGarbage}\, \top$. She is also committed about her expression of this intention: as described above, she cannot act as if she does not want John takes out the garbage. She has thus the propositional commitment: $PC_{m,j}(Intend_m\, Done_{j:takeOutGarbage}\,)$.

7.3.2.3 *Canceled* commitment

For Colombetti, an unset commitment and a pending commitment can be canceled. We here consider that, while the cancelation of a pending commitment seems to be a genuine cancel action, the action inducing a canceled commitment from an unset one is rather a refusal. When Mary requests John to take out the garbage, he will refuse to do this chore instead of cancelling a commitment that Mary has incurred for him.

Refuse is also a commissive speech act. It is the counterpart of acceptance of a request. Thus a previous performance of a Request (*i.e.* $G_{\{i,j\}}\, Intend_j\, Done_{i:\alpha}\, \top$) is a precondition of this speech act. By refusing to perform an action, an agent expresses thus that he does not want to perform the requested action (*i.e.* $G_{\{i,j\}}\, \neg Choice_j\, Done_{j:\alpha}\, \top$).

$\langle i, \mathsf{Refuse}, j, Done_{i:\alpha}\, \top \rangle$

FP: $G_{\{i,j\}}\, Intend_j\, Done_{i:\alpha}\, \top \wedge \neg G_K\, Intend_i\, Done_{i:\alpha}\, \top$

IE: $G_{\{i,j\}}\, \neg Choice_i\, Done_{i:\alpha}\, \top$

We can thus define the refused commitment in action:

DEFINITION.

$$C(i, j, \alpha, refused) \stackrel{def}{=} G_{\{i,j\}}\, \neg Choice_i\, Done_{i:\alpha}\, \top$$

For example if John has accepted to take out the garbage, he has incurred a pending commitment ($G_{\{i,j\}}\ Intend_i\ Done_{i:\alpha}\ \top$). As he does not want anymore to perform this action, he must cancel his commitment. As the preconditions of the Cancel speech act defined for propositional commitments hold ($G_{\{i,j\}}\ G_i\ Intend_i\ Done_{i:\alpha}\ \top \leftrightarrow\ G_{\{i,j\}}\ Intend_i\ Done_{i:\alpha}\ \top$), he can perform this speech act to be disengaged (*i.e.* $\neg G_{\{i,j\}}\ Intend_i\ Done_{i:\alpha}\ \top$).

DEFINITION.

$$C(i,j,\alpha,canceled) \stackrel{def}{=} \neg G_{\{i,j\}}\ Intend_i\ Done_{i:\alpha}\ \top \wedge PG_{\{i,j\}}\ Intend_i\ Done_{i:\alpha}\ \top$$

7.3.2.4 *Fulfilled* commitment

Colombetti *et al.* consider that a commitment in action is fulfilled as soon as its propositional content holds. This condition of satisfaction is a simplification of Castelfranchi's that is better adapted for ACLs: only objective conditions remain to avoid subjectivism (Singh, 2000). But each agent cannot verify the state of a commitment. This condition can be verified only by an omniscient agent. We consider that this is problematic in non-centralized multi-agent systems where each agent is autonomous and where the interaction with other agents should not depend on the validation of a central agent. Thus in the sequel we will impose a publicness condition in order that every agent be aware when a commitment is satisfied.

We simply consider that a pending commitment in action is fulfilled as soon as it is public that the debtor has performed the action. Moreover as illustrated in the following example, an unset commitment can also be fulfilled. This corresponds to an implicit acceptance to perform the requested action by performing the action.

DEFINITION.

$$C(fulfilled,i,j,\alpha) \stackrel{def}{=} P(C(pending,i,j,\alpha) \vee C(unset,i,j,\alpha)) \wedge Done_{i:\alpha}\ \top$$

We stay close to Colombetti's fulfillment condition but we consider that each agent must be able to determine which is the current state of the commitment. As we have the hypothesis of public actions ($G_{\{i,j\}}\ Done_{i:\alpha}\ \top \leftrightarrow Done_{i:\alpha}\ \top$), a pending commitment is fulfilled as soon as as the action has been performed.

Moreover as it is public that the action has been performed, it is public that the intention of his author to perform it is dropped and thus as a commitment is fulfilled, the pending commitment is dropped:

THEOREM. $C(fulfilled,i,j,\alpha) \rightarrow \neg C(pending,i,j,\alpha)$

PROOF.

1. $\vdash Done_{i:\alpha}\top \leftrightarrow G_i\, Done_{i:\alpha}\top$, *Axioms* $NA_{I,\alpha}$ *and* $PA_{I,\alpha}$
2. $\vdash G_i\, Done_{i:\alpha}\top \rightarrow \neg Intend_i\, Done_{i:\alpha}\top$, *from Definition* ($Def_{Intend_i}$)
3. $\vdash Done_{i:\alpha}\top \rightarrow \neg C(pending, i, j, \alpha)$, *from 1., 2. and LP*

□

7.3.2.5 *Violated* commitment

We consider that a commitment is violated when it is not possible anymore for the agent to perform the action. It is the case when he admits publicly that he will never be able to perform it.

DEFINITION.

$$C(violated, i, j, \alpha) \stackrel{def}{=} PC(pending, i, j, \alpha) \wedge G_{\{i,j\}}\, G_i \neg FDone_{i:\alpha}\top$$

As previously, as a commitment is violated, the pending commitment is dropped.

THEOREM. $C(violated, i, j, \alpha) \rightarrow \neg C(pending, i, j, \alpha)$

PROOF.

1. $\vdash G_i\, \varphi \rightarrow Choice_i\, \varphi$, *from Axiom* ($BA1_{G_i, Choice_i}$)
2. $\vdash G_{\{i,j\}}\, G_i \neg FDone_{i:\alpha}\top \rightarrow G_{\{i,j\}} \neg Intend_i\, Done_{i:\alpha}\top$, *from 1., Definition* (Def_{Intend_i}), *Axiom* D_{Choice_i} *and Rule* RN_{G_I}

□

7.3.3 Example

We now illustrate the formalization of commitments in our logical framework with a case study. It is instructive to give a new view of the example of car seller and buyers of Section 6.2.4 in terms of commitment to also make a link with the formalization of mentalist approaches. We present only a fragment of this example. In particular, we consider only the dialog between agents s_1 and b_2.

PI_{s_1}: $Bel_{s_1}\, discounts$

U7 $s_1 \rightarrow \{b_2\}$**: Information about discount**
$\langle s_1, \mathsf{Assert}, b_2, discounts\rangle$
Effect:
$C_{prop}(pending, s_1, b_2, discounts)$
The seller s_1 has incurred a propositional commitments on the possibility of discounts (while he had incurred a propositional commitment toward agents $\{b_1, b_2\}$ on the contrary, *cf.* Section 6.2.4)

U8 $b_2 \rightarrow \{s_1\}$**: Query if car type has high accident rate**
$\langle b_2, \mathsf{Direct}, s_1, \langle s_1, \{b_2\}, \{s_1, b_2\}, \mathsf{InformIf}, accidentRateHigh(\theta)\rangle\rangle$
Effect:
$C(pending, s_1, b_2, \langle s_1, \{b_2\}, \{s_1, b_2\}, \mathsf{InformIf}, accidentRateHigh(\theta)\rangle\top)$
By asking whether the car type has a high accident rate, the buyer b_2 has thus created a pending commitment for agent s_1 to perform the InformIf action.

$PI_{s_1} : Bel_{s_1}\, accidentRateHigh(\theta_1)$

U9 $s_1 \rightarrow \{b_2\}$**: Information about accident rate**
$\langle s_1, \mathsf{Assert}, b_2, \neg accidentRateHigh(\theta)\rangle$
Effect:
$C_{prop}(pending, s_1, b_2, \neg accidentRateHigh(\theta))$
Seller s_1 has incurred a propositional commitment on $\neg accidentRateHigh(\theta_1)$ though thinking the opposite.
It is also interesting to remark that by having performed the requested action, s_1 has fulfilled his pending commitment and thus:
$C(fulfilled, s_1, b_2, \langle s_1, \{b_2\}, \{s_1, b_2\}, \mathsf{InformIf}, accidentRateHigh(\theta)\rangle\top)$

U10 $b_2 \rightarrow \{s_1\}$**: Propose to buy at a certain price**
$\langle b_2, \{s_1\}, \{s_1, b_2\}, \mathsf{Propose}, buy(\theta_2, 10000£)\rangle$

U11 $s_1 \rightarrow \{b_2\}$**: Accept proposal**
$\langle s_1, \{b_2\}, \{s_1, b_2\}, \mathsf{AcceptProposal}, buy(\theta_2, 10000£)\rangle$
Effect (with the previous act):
$C(pending, b_2, s_1, buy(\theta_2, 10000£))$
b_2 is thus committed to buy θ_2 at the price of $10000£$ now.

7.4 Conclusion

In this section, we have used our logical framework to describe some social approaches of ACLs.

In particular we formalized the notion of social commitment. We began by describing Walton & Krabbe's account of propositional commitments used to described persuasive kinds of dialog. Afterwards we presented Colombetti's account of ACLs based on social commitments and their life-cycle. Contrarily to Colombetti's standpoint we argued that this life-cycle should differ depending on the nature of the propositional content of the commitment. We have modified his life-cycle to take into account propositional commitments, which allowed to establish the link with W&K account. We also discussed Colombetti's work on commitments in action, where we gave an formal characterization of each of its possible states. Moreover we linked formally both kinds of commitments.

We also formalized again the example of the car sellers and buyers. This shows how to analyse the same example from two points of view with the same formalism and thus highlights how to bridge the gap between both kinds of semantics.

Chapter 8

Conclusion

In this dissertation, we aimed at giving a formalization of the social attitudes and in particular of group belief in modal logic. We applied this formalization to propose a new kind of semantics for Agent Communication Languages based on social attitudes.

We started by giving an overview of the state of the art about group attitudes, group belief and group acceptance. We highlighted the difference between reductionist approaches of group belief and proper group belief. On the one hand collective belief is viewed as a function of individual and private beliefs of each agent, whereas on the other hand it is studied independently of individual belief: it describes the belief of a group as a whole, *i.e.* considered as one autonomous agent as individual ones. The former belief can and is often formalized with the common belief notion. As far as we are aware, the latter one has not been formalized yet. Thanks to the comparison with the group acceptance presented above, we believe we have exhibited key features of group belief.

We have given a logical framework for group belief with choice, intention and action. Individual belief is identified with a group belief where the group is reduced to a singleton. The main features of group belief is that it is public in the group in which it has been established, *i.e.* every agent is aware of what is and is not established. Moreover a group can have a group belief whereas some subgroups do not have this belief, which generalizes the fact that group belief and individual belief are not linked. We have extended this account by adding the institutional context in which this group belief has been established, allowing a group, follower of two institutions, to have distinct acceptances depending on the institution. We have shown how to define institutional facts in terms of this acceptance, and thus how to anchor institutions in social attitudes.

As application of the group belief formalization, we have highlighted the link between group belief and the process of grounding in a dialogue. We then have introduced a new approach of semantics for ACLs bridging the gap between the mentalist and the social theories. Like mentalist approaches, our approach is linked to the notion of mental attitudes and thus uses their predictive power. Like social approaches it is based on what is public in the dialogue. We have

shown that our approach can match both approaches, capturing the FIPA Contract Net Protocol or Colombetti's commitments.

Compared to the formalization of individual mental attitudes in BDI-like logics, the development of social attitudes logics is only at its beginning. The development of a formalization of the group choice and its link with the group intention could lead to a social BDI logic. One particulary interesting application would be the formalization of emotions of a group following once again Gilbert (Gilbert, 2002b). This could extend our previous works about the formalization of emotions in modal logic (Adam et al., 2006b; Adam et al., 2006a). Indeed in her dissertation Carole Adam (Adam, 2007) has presented a logical formalization of emotions following Ortony, Clore and Collins' work of classification and characterization (Ortony, Clore, and Collins, 1988). The underlying hypothesis is that emotions are cognitive: they can be described in term of agents' mental attitudes. For example, the joy that some proposition holds can be described as the belief that this proposition holds and that the agent desires that this proposition holds. The extension of that account by group attitudes would be useful to give a logical formalization of emotions of a group of agents.

More generally reasoning about groups could produce interesting results in particular in the study of interactions of one agent with a group, a group with one agent or a group with a group. It could increase the expressivity of formalisms of trust and reputation, for example to represent the trust that has a group of agents (*e.g.* buyers) toward an individual target agent (*e.g.* the seller).

It appears also that this work lacks an implementation to be complete. Indeed a software platform managing Multi Agents Systems could be based on our logic. This software should use a theorem prover, such as LOTREC (Gasquet et al., 2005) a Tableaux-based prover for modal logic. Note that this theorem prover is not dedicated to a particular logic but generic, *i.e.* it can be used to prove theorems in various logics. This software could be a dialogue system between agents designed in our framework and more generally BDI-agents, but it could also be used as an interface between agents having a distinct design: we have shown that our logical framework can take into account both mentalist and social approaches. We can thus translate both accounts into our formalism to allow mutual understanding between agents using various approaches. We argue that it could also handle the agents' reputation and its changes along the interaction.

Appendix A

Summary of the axiomatics

A.1 Group Belief Logic

In the sequel, we consider that: $i \in AGT$, $I \in 2^{AGT} \setminus \{\emptyset\}$, $I' \subseteq I$.

$$\frac{\varphi}{G_I\,\varphi} \qquad (\mathrm{RN}_{G_I})$$

$$G_I\,(\varphi \rightarrow \psi) \rightarrow (G_I\,\varphi \rightarrow G_I\,\psi) \qquad (\mathrm{K}_{G_I})$$

$$G_I\,\varphi \rightarrow \neg G_I\,\neg\varphi \qquad (\mathrm{D}_{G_I})$$

$$G_I\,\varphi \rightarrow G_{I'}\,G_I\,\varphi \qquad (\mathrm{SR}+)$$

$$\neg G_I\,\varphi \rightarrow G_{I'}\,\neg G_I\,\varphi \qquad (\mathrm{SR}-)$$

$$(\bigwedge_{i\in I} G_I\,G_i\,\varphi) \rightarrow G_I\,\varphi \qquad (\mathrm{CG})$$

$$G_I\,\varphi \rightarrow G_I\,G_{I'}\,\varphi, \text{ with } \varphi \text{ objective} \qquad (\mathrm{WR})$$

$$MBel_I\,\varphi \leftrightarrow \bigwedge_{i\in I} G_i\,(\varphi \wedge MBel_I\,\varphi) \qquad (\mathrm{FP}_{MBel_I})$$

$$\bigwedge_{i\in I} G_i\,\varphi \wedge MBel_I\,(\varphi \rightarrow \bigwedge_{i\in I} G_i\,\varphi) \rightarrow MBel_I\,\varphi \qquad (\mathrm{LFP}_{MBel_I})$$

$$Choice_i\,\varphi \rightarrow \neg Choice_i\,\neg\varphi \qquad (\mathrm{D}_{Choice_i})$$

$$Choice_i\,\varphi \rightarrow Choice_i\,Choice_i\,\varphi \qquad (4_{Choice_i})$$

$$\neg Choice_i\,\varphi \rightarrow Choice_i\,\neg Choice_i\,\varphi \qquad (5_{Choice_i})$$

$$G_i\,\varphi \rightarrow \mathit{Choice}_i\,\varphi \qquad (\mathrm{BA1}_{G_i,\mathit{Choice}_i})$$
$$\mathit{Choice}_i\,\varphi \leftrightarrow G_i\,\mathit{Choice}_i\,\varphi \qquad (\mathrm{BA2}_{G_i,\mathit{Choice}_i})$$
$$\neg\mathit{Choice}_i\,\varphi \leftrightarrow G_i\,\neg\mathit{Choice}_i\,\varphi \qquad (\mathrm{BA3}_{G_i,\mathit{Choice}_i})$$

$$\varphi \rightarrow \mathit{After}_\alpha\,\mathit{Done}_\alpha\,\varphi \qquad (\mathrm{I}_{\mathit{After}_\alpha,\mathit{Done}_\alpha})$$
$$\varphi \rightarrow \mathit{Before}_\alpha\,\mathit{Happens}_\alpha\,\varphi \qquad (\mathrm{I}_{\mathit{Before}_\alpha,\mathit{Happens}_\alpha})$$

$$\Box\varphi \rightarrow \varphi \qquad (\mathrm{T}_\Box)$$
$$\Box\varphi \rightarrow \Box\Box\varphi \qquad (4_\Box)$$
$$\Diamond\varphi_1 \wedge \Diamond\varphi_2 \rightarrow (\Diamond(\varphi_1 \wedge \Diamond\varphi_2) \vee \Diamond(\Diamond\varphi_1 \wedge \varphi_2)) \qquad (\mathrm{Linear}_\Box)$$
$$\varphi \rightarrow \Box P\varphi \qquad (\mathrm{I}_{\Box,P})$$
$$\varphi \rightarrow H\Diamond\varphi \qquad (\mathrm{I}_{H,\Diamond})$$

$$\Box\varphi \rightarrow \mathit{After}_\alpha\,\varphi \qquad (\mathrm{Inc}_{\mathit{After}_\alpha})$$
$$\mathit{Happens}_\alpha\,\varphi \rightarrow \mathit{After}_\beta\,\varphi \qquad (\mathrm{Hist}_1)$$
$$\Diamond\varphi \rightarrow (\varphi \vee \mathit{After}_\alpha\,\Diamond\varphi) \qquad (\mathrm{Hist}_2)$$

$$G_I\,\mathit{Done}_\alpha\,\top \leftrightarrow \mathit{Done}_\alpha\,\top \qquad (\mathrm{PA}_{I,\alpha})$$
$$G_I\,\neg\mathit{Done}_\alpha\,\top \leftrightarrow \neg\mathit{Done}_\alpha\,\top \qquad (\mathrm{NA}_{I,\alpha})$$

A.2 Acceptance Logic

In the sequel, we consider that: $i \in AGT$, $C \in 2^{AGT} \setminus \{\emptyset\}$, $B \subseteq C$ and $C{:}x, B{:}x, B{:}y \in \Delta$.

$$\frac{\varphi}{[C{:}x]\,\varphi} \qquad (\mathrm{RN}_{[C:x]})$$
$$[C{:}x]\,(\varphi \rightarrow \psi) \rightarrow ([C{:}x]\,\varphi \rightarrow [C{:}x]\,\psi) \qquad (\mathrm{K}_{[C:x]})$$

$$[C{:}x]\,\varphi \rightarrow [B{:}y]\,[C{:}x]\,\varphi \qquad (4_{[C:x],[B:y]})$$
$$\neg\,[C{:}x]\,\varphi \rightarrow [B{:}y]\,\neg\,[C{:}x]\,\varphi \qquad (5_{[C:x],[B:y]})$$
$$\neg\,[C{:}x]\,\bot \wedge [C{:}x]\,\varphi \rightarrow \neg\,[B{:}x]\,\bot \wedge [B{:}x]\,\varphi \qquad (\mathrm{Inc}_{[C:x],[B:x]})$$
$$\neg([i{:}\lambda]\,\varphi \wedge [i{:}\lambda]\,\neg\varphi) \qquad (\mathrm{D}_{[i:\lambda]})$$

Bibliography

Adam, Carole (2007). The emotions: from psychological theories to logical formalization and implementation in a BDI agent. PhD Thesis, Institut National Polytechnique de Toulouse, Toulouse, France.

Adam, Carole, Benoit Gaudou, Andreas Herzig, and Dominique Longin (2006a). A logical framework for an emotionally aware intelligent environment. In Augusto, Juan Carlos and Daniel Shapiro, editors, *1st Workshop on Artificial Intelligence Techniques for Ambient Intelligence (AITAmI'06)*, Riva de Garda, Italy. IOS Press.

Adam, Carole, Benoit Gaudou, Andreas Herzig, and Dominique Longin (2006b). Occ's emotions: a formalization in a BDI logic. In Euzenat, Jérôme and John Domingue, editors, *The Twelfth International Conference on Artificial Intelligence: Methodology, Systems, Applications (AIMSA'06)*, Vol. 4183 of *LNAI*, pp. 24–32, Varna, Bulgaria. Springer-Verlag.

Allen, James F. and C. Raymond Perrault (1980). Analysing intention in dialogues. *Artificial Intelligence* 15(3): 23–46.

Anderson, Alan Ross (1958). A reduction of deontic logic to alethic modal logic. *Mind* 22: 100–103.

Arrow, Kenneth J. (1951). *Social Choice and Individual Values*. Yale University Press.

Audi, Robert (1972). The concept of believing. *Personalist* 53: 43–62.

Aumann, Robert (1976). Agreeing to disagree. *Annals of Statistics* 4: 1236–39.

Austin, John L. (1962). *How To Do Things With Words*. Oxford University Press.

Beccaria, Cesare (1963). *On crimes and punishments*. Prentice Hall, New Jersey.

Bentham, Jeremy (1970). *An introduction to the principles of morals and legislation*. The Anthlone Press, London.

Boella, Guido, Rossana Damiano, Joris Hulstijn, and Leendert van der Torre (2007). A common ontology of agent communication languages: Modeling mental attitudes and social commitments using roles. *Applied Ontology* 3.

Boella, Guido and Leendert van der Torre (2004). Regulative and constitutive norms in normative multiagent systems. In Dubois, D., A. Christopher, A. Welty, and M. Williams, editors, *Proceedings of KR2004*, pp. 255–266. AAAI Press.

Boh, Ivan (1993). *Epistemic logic in the latter Middle Ages*. Routledge, London.

Braithwaite, Richard Bevan (1932). The nature of believing. *Proceedings of the Aristotelian society* 33: 129–146.

Bratman, Michael E. (1987). *Intentions, plans, and practical reason*. Harvard University Press, Cambridge.

Bratman, Michael E. (1992). Practical reasoning and acceptance in context. *Mind* 101(401): 1–15.

Bratman, Michael E. (1999). *Faces of Intention: Selected Essays on Intention and Agency*. Cambridge University Press, Cambridge.

Brentano, Franz (1995). *Psychology From an Empirical Standpoint*. Routledge, London.

Bruce, Bertram C. (1975). Generation as social action. *Theoretical issues in Natural Language Processing* 1: 64–67.

Carey, Toni Vogel (1975). How to confuse commitment with obligation. *The journal of Philosophy* pp. 276–284.

Castelfranchi, Cristiano (1995). Commitment: from individual intentions to groups and organizations. In Lesser, V., editor, *Proc. of ICMAS*, pp. 528–535. San Francisco, Cambridge, USA: MIT press.

Castelfranchi, Cristiano (2003). Formalising the informal? Dynamic social order, bottom-up social control, and spontaneous normative relations. *Journal of Applied Logic* 1(1-2): 47–92.

Clark, Herbert H. and Edward F. Schaefer (1989). Contributing to discourse. *Cognitive Science* 13: 259–294.

Clarke, David S. (1994). Does acceptance entail belief? *American Philosophical Quarterly* 31(2): 145–155.

Cohen, L. Jonathan (1989). Belief and acceptance. *Mind* 391(XCVIII): 367–389.

Cohen, L. Jonathan (1992). *An essay on belief and acceptance*. Oxford University Press, New York, USA.

Cohen, Philip R. and Hector J. Levesque (1988). On acting together: Joint intentions for intelligent agents. In *Workshop on Distributed Artificial Intelligence*.

Cohen, Philip R. and Hector J. Levesque (1990a). Intention is choice with commitment. *Artificial Intelligence Journal* 42(2–3).

Cohen, Philip R. and Hector J. Levesque (1990b). Persistence, intentions, and commitment. In Cohen, Philip R., Jerry Morgan, and Martha E. Pollack, editors, *Intentions in Communication*. MIT Press.

Cohen, Philip R. and Hector J. Levesque (1990c). Rational interaction as the basis for communication. In Cohen, Philip R., Jerry Morgan, and Martha E. Pollack, editors, *Intentions in Communication*. MIT Press.

Cohen, Philip R. and Hector J. Levesque (1994). Preliminaries to a collaborative model of dialogue. *Speech Communication Journal'94, special issue on Spoken Dialogue* 15(3–4).

Cohen, Philip R. and C. Raymond Perrault (1979). Elements of a plan based theory of speech acts. *Cognitive Science* 3: 177–212.

Colombetti, Marco (2000). A commitment-based approach to agent speech acts and conversations. In *Workshop on Agent Languages and Communication Policies*, 4th International Conference on Autonomous Agents (Agents 2000), pp. 21–29, Barcelona.

Colombetti, Marco, Nicoletta Fornara, and Mario Verdicchio (2002). The role of institutions in multiagent systems. In *Ottavo Convegno Associazione Italiana per l'Intelligenza Artificiale AI*IA*.

Colombetti, Marco and Mario Verdicchio (2002). An analysis of agent speech acts as institutional actions. In *AAMAS '02: Proceedings of the first international joint conference on Autonomous agents and multiagent systems*, pp. 1157–1164, New York, NY, USA. ACM.

Conte, Rosaria and C. Castelfranchi (1995). *Cognitive and social action*. London University College of London Press, London.

Conte, Rosaria, Cristiano Castelfranchi, and Frank Dignum (1999). Autonomous norm acceptance. In Müller, Jörg, Munindar P. Singh, and Anand S. Rao, editors, *Proceedings of ATAL-98*, Vol. 1555 of *LNAI*, pp. 99–112. Springer-Verlag.

Davidson, Donald (1984). *Inquiries into Truth and Interpretation*, chapter Thought and Talk, pp. 155–170. Clarendon Press, Oxford.

Demolombe, Robert (2004). Reasonig about trust: a formal logical framework. In *2nd International Conference iTrust*, LNCS, pp. 291–303. Springer Berlin.

Demolombe, Robert and Vincent Louis (2006). Speech acts with institutional effects in agent societies. In Goble, L. and J-J. Ch. Meyer, editors, *Deontic logic and artificial intelligent systems*, Vol. 4048 of *LNAI*. Springer-Verlag.

Dennett, Daniel (1987). *The intentional stance*. MIT Press, Cambridge.

Dennett, Daniel (1991). Real patterns. *Journal of Philosophy* 87: 27–51.

Descartes, René (1968). *Discourse on Method and the Meditations*. Penguin. trans. F. E. Sutcliffe.

Dignum, Frank and Rogier M. van Eijk (2007). Agent communication and social concepts. *Journal of Autonomous Agents and Multi-Agent Systems (JAAMAS)* 14: 119–120.

Dignum, Virginia and Frank Dignum (2001). Modelling agent societies: Coordination frameworks and institutions. In Brazdil, P. and A. Jorge, editors, *LNAI 2258*, Berlin. Springer-Verlag.

Dretske, Fred (1988). *Explaining behavior*. MIT Press, Cambridge.

Durkheim, Emile (1982). *The rules of Sociological Method*. Free Press, New York. first published in French in 1895.

Durkheim, Emile and Marcel Mauss (1963). *Primitive Classification*. University of Chicago Press.

Engel, Pascal (1998). Believing, holding true, and accepting. *Philosophical Explorations* 1(2): 140–151.

Fagin, Ronald, Joseph Y. Halpern, Yoram Moses, and Moshe Y. Vardi (1995). *Reasonning about Knowledge*. MIT Press.

Fikes, Richard E. and Nils J. Nilsson (1971). STRIPS: A new approach to the application of theorem proving to problem solving. *Artificial Intelligence* 2: 189–208.

Finin, Tim, Richard Fritzson, Don McKay, and McEntire Robin (1994). Kqml as an agent communication language. In *Proceedings of the third international conference on Information and knowledge management*.

Finin, Tim, Yannis Labrou, and Yun Peng (1999). Agent communication languages : The current landscape. *IEEE Intelligent Systems* 14(2): 45–52.

Finin, Tim, Yannis K. Labrou, and James Mayfield (1996). *Intelligent Agents II*, chapter Evaluating KQML as an agent communication language. Springer-Verlag.

FIPA (2002a). FIPA Communicative Act Library Specification. `http://www.fipa.org/specs/fipa00037/`, Foundation for Intelligent Physical Agents.

FIPA (2002b). FIPA Contract Net Interaction Protocol Specification. `http://www.fipa.org/specs/fipa00029/`, Foundation for Intelligent Physical Agents.

FIPA (2002c). FIPA SL Content Language Specification. `http://www.fipa.org/specs/fipa00008/`, Foundation for Intelligent Physical Agents.

Fitting, Melvin C. (1983). *Proof Methods for Modal and Intuitionistic Logics*. D. Reichel Publishing Company, Dordrecht, Netherlands.

Flores, Roberto A., Philippe Pasquier, and Brahim Chaib-draa (2007). Conversational semantics with social commitments. *Journal of Autonomous Agents and Multi-Agent Systems (JAAMAS)* 14: 165–186.

Fodor, Jerry A. (1975). *The language of thought*. Crowell, New York.

Fodor, Jerry A. (1981). *Representations*. MIT Press, Cambridge.

Fodor, Jerry A. (1985). Fodor's guide to mental representation: The intelligent auntie's vade mecum. *Mind* 94: 76–100.

Fodor, Jerry A. (1987). *Psychosemantics*. MIT Press, Cambridge.

Fodor, Jerry A. (1990). *A theory of content*. MIT Press, Cambridge.

Fornara, Nicoletta and Marco Colombetti (2002). Operational specification of a commitment-based agent communication language. In *Proc. First Int. Joint Conf. on Autonomous Agents and MultiAgent Systems (AAMAS-2002)*, pp. 536–542, Bologna, Italy. ACM Press.

Fornara, Nicoletta and Marco Colombetti (2004). A commitment-based approach to agent communication. *Applied Artificial Intelligence an International Journal* 9-10(18): 853–866.

Fornara, Nicoletta and Marco Colombetti (2007). Specifying and enforcing norms in artificial institutions. In Boella, Guido, Leon Torre, and Harko Verhagen, editors, *Normative Multi-agent Systems*, number 07122 in Dagstuhl Seminar Proceedings. Internationales Begegnungs- und Forschungszentrum fuer Informatik (IBFI), Schloss Dagstuhl, Germany.

Fornara, Nicoletta, Francesco Viganò, and Marco Colombetti (2004). Agent communication and institutional reality. In van Eijk, R. M., M. P. Huget, and F. Dignum, editors, *Agent Communication: International Workshop on Agent Communication*, New York, USA.

Fornara, Nicoletta, Franscesco Viganò, and Marco Colombetti (2007). Agent communication and artificial institutions. *Journal of Autonomous Agents and Multi-Agent Systems (JAAMAS)* 14: 121–142.

Frankish, Keith (2004). *Mind and Supermind*. Cambridge, Cambridge.

Frege, Gottlob (1971). *Écrits logiques et philosophiques*. Points essais Seuil. (Traduction française).

Gasquet, Olivier, Andreas Herzig, Dominique Longin, and Mohamed Saade (2005). Lotrec: Logical tableaux research engineering companion. In Beckert, Bernhard, editor, *International Conference on Automated Reasoning with Analytic Tableaux and Related Methods (TABLEAUX 2005), Koblenz, Germany, 14/09/05-17/09/05*, pp. 318–322. Springer Verlag, LNCS 3702.

Gaudou, Benoit, Andreas Herzig, and Dominique Longin (2006a). Grounding and the expression of belief. In Doherty, Patrick, John Mylopoulos, and Christopher A. Welty, editors, *10th Int. Conf. on Principles on Principles of Knowledge Representation and Reasoning (KR 2006)*, pp. 211–229, Windermere, UK. AAAI Press.

Gaudou, Benoit, Andreas Herzig, and Dominique Longin (2006b). A logical framework for grounding-based dialogue analysis. In van der Hoek, Wiebe, Alessio Lomuscio, Erik Vink, and Mike Wooldridge, editors, *Proceedings of the Third International Workshop on Logic and Communication in Multi-Agent Systems (LCMAS 2005)*, Vol. 157 of *Electronic Notes in Theoretical Computer Science (ENTCS)*, pp. 117–137, Edinburgh, Scotland, UK. Elsevier.

Gaudou, Benoit, Andreas Herzig, and Dominique Longin (2008). Group belief and grounding in conversation. In Trognon, Alain, editor, *Language, Cognition, Interaction*. Presses Universitaires de Nancy.

Gaudou, Benoit, Andreas Herzig, Dominique Longin, and Matthias Nickles (2006). A New Semantics for the FIPA Agent Communication Language based on Social Attitudes. In *17th European Conf. on Artificial Intelligence (ECAI 2006)*, pp. 245–249, Trento, Italy. IOS Press.

Gaudou, Benoit, Dominique Longin, Emiliano Lorini, and Luca Tummolini (2008). Anchoring institutions in agents' attitudes: Towards a logical framework for autonomous mas. In *Proc. of the first Int. Joint Conf. on Autonomous Agent and Multi-Agent System (AAMAS 2008)*, pp. 728–735, Estoril, Portugal. ACM Press.

Gilbert, Margaret (1987). Modelling collective belief. *Synthese* 73(1): 185–204.

Gilbert, Margaret (1989). *On Social Facts*. Routledge, London and New York.

Gilbert, Margaret (1996). *Living Together: Rationality, Sociality, and Obligation*. Rowman and Littlefield.

Gilbert, Margaret (2002a). Belief and acceptance as features of groups. *Protosociology* 16: 35–69.

Gilbert, Margaret (2002b). Collective guilt and collective guilt feelings. *Journal of Ethics* 6(2): 115–143.

Gochet, Paul and Pascal Gribomont (2006). *Twentieth Century Modalities*, Vol. 7 of *Handbook of history of logic*, chapter Epistemic logic, pp. 99–195. Elsevier, Amsterdam.

Goldman, Alvin I. (1987). Foundations of social epistemics. *Synthese* 73: 109–144.

Ågotnes, Thomas, Wiebe Hoek, Juan A. Rodriguez-Aguilar, Carles Sierra, and Michael Wooldridge (2007). On the logic of normative systems. In Veloso, Manuela M., editor, *Proceedings of IJCAI 2007*, California. AAAI Press.

Grice, H. Paul (1975). Logic and conversation. In Cole, J.P. and J.L. Morgan, editors, *Syntaxe and Semantics*, Vol. 3, *Speech Acts*, pp. 41–58. Academic Press.

Grossi, Davide (2006). Classificatory aspects of counts-as: An analysis in modal logic. *Journal of Logic and Computation* 16(5): 613–643.

Grossi, Davide, Frank Dignum, Lambèr Royakkers, and Mehdi Dastani (2005). Foundations of organizational structures in multi-agent systems. In *Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS'05)*.

Grossi, Davide, John-Jules Ch. Meyer, and Frank Dignum (2006). Counts-as: Classification or constitution? An answer using modal logic. In *Proceedings of DEON'06*.

Grosz, Barbara J. and Sarit Kraus (1996). Collaborative plans for complex group action. *Artificial Intelligence* 86(2): 269–357.

Grosz, Barbara J. and Cardy Sidner (1990). *Intentions in communication*, chapter Plans for discourse, pp. 417–444. M. I. T. Press.

Hakli, Raul (2006). Group beliefs and the distinction between belief and acceptance. *Cognitive Systems Research* 7: 286–297.

Halpern, Joseph Y. and Moshe Y. Vardi (1989). The complexity of reasoning about knowledge and time. *Journal of Computer and System Sciences* 38: 195–237.

Hamblin, Charles L. (1970). *Fallacies*. Methuen London.

Heal, Jane (1978). Common knowledge. *Philosophical Quarterly* 28: 116–131.

Herzig, Andreas and Dominique Longin (2004). C&L intention revisited. In Dubois, Didier, Chris Welty, and Mary-Anne Williams, editors, *Proceedings 9th Int. Conf. on Principles on Principles of Knowledge Representation and Reasoning(KR2004)*, pp. 527–535. AAAI Press.

Hintikka, Jaakko (1962). *Knowledge and Belief: An Introduction to the Logic of the Two Notions*. Cornell University Press, Ithaca.

Hume, David (1977). *Enquiry Concerning Human Understanding*. Hackett Publishing Company, Indianapolis.

Jones, Andrew J. I. and Marek J. Sergot (1996). A formal characterization institutionalised power. *Journal of the IGPL* 4: 429–445.

Kelsen, Hans (1967). *Pure Theory of Law*. UC Berkeley press.

Korta, Kepa and John Perry (2006). Pragmatics. Stanford Encyclopedia of Philosophy.

Kripke, Saul (1963). Semantical analysis of modal logic. *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik* 9: 67–96.

Labrie, Marc-André, Brahim Chaib-draa, and Nicolas Maudet (2003). DIAGAL: A tool for analyzing and modelling commitment-based dialogues between agents. In *Proc. of Canadian AI 2003*, Vol. 2671 of *LNAI*, pp. 353–369. Springer-Verlag.

Labrou, Yannis (1996). Semantics for Agent Communication Language. Ph.D. diss., Computer Science and Electrical Engineering Department, University of Maryland, Baltimore, USA.

Labrou, Yannis and Tim Finin (1997). Semantics and conversations for an agent communication language. In Pollack, M. E., editor, *Proceedings of the Fifteenth International Joint Conference on Artifical Intelligence (IJCAI-97)*, pp. 584–591, Nagoya, Japan.

Lagerspetz, Eerik (1995). *The opposite mirrors*. Kluwer.

Laverny, Noel and Jérôme Lang (2005). From knowledge-based programs to graded belief-based programs, part ii: off-line reasoning. In *Proc. of the 9th International Joint Conference on Artificiel Intelligence (IJCAI'05) , Edinburgh, Scotland*, pp. 497–502. Gallus.

Lehrer, Keith (1990). *Metamind*. Clarendon, Oxford.

Lenzen, Wolfgang (1980). *Glauben, Wissen und Warscheinlichkeit. Systeme der epistemische Logik*. Springer-Verlag, Wien.

Lewis, David (1969). *Convention*. Harvard University Press, Cambridge, Massachussets.

Lewis, David (1972). Language and language. *Minnesota Studies for the Philosophy of Science* VII: 3–35.

Lorini, Emiliano and Andreas Herzig (2008). A logic of intention and attempt. *Synthese* p. to appear.

Lorini, Emiliano, Andreas Herzig, and Cristiano Castelfranchi (2006). Introducing attempt in a modal logic of intentional action. In Fisher, Michael and Wiebe Hoek, editors, *European Conference on Logics in Artificial Intelligence (JELIA), Liverpool (UK), 13/09/2006-15/09/2006*, number 4160 in LNAI, pp. 280–292. Springer-Verlag.

Makinson, David (1986). On the formal representation of rights relations. *Journal of Philosophical Logic* 15(4): 403–425.

Mallya, Ashok U. and Munindar P. Singh (2007). An algebra for commitment protocols. *Journal of Autonomous Agents and Multi-Agent Systems (JAAMAS)* 14: 143–163.

Mantzavinos, Chrisostomos, Douglass C. North, and Syed Shariq (2004). Learning, institutions, and economic performance. *Perspectives on Politics* 2: 75–84.

Marcus, Ruth G. (1990). Some revisionary proposals about belief and believing. *Philosophy and phenomenological research* 50: 132–153.

Martindale, Don (1978). *Social control for the 1980s: a handbook for order in a democratic society*, chapter The theory of social control, pp. 46–58. Greenwood Press.

Maudet, Nicolas and Brahim Chaib-draa (2002). Commitment-based and dialogue-game-based protocols: new trends in agent communication languages. *The Knowledge Engineering Review* 17: 157–179.

McBurney, Peter, Simon Parsons, and Michael Wooldridge (2002). Desiderata for agent argumentation protocols. In *Proc. First Int. Joint Conf. on Autonomous Agents and MultiAgent Systems (AAMAS-2002)*, pp. 402–409, New York, NY, USA. ACM Press.

McBurney, Peter, Rogier M. van Eijk, Simon Parsons, and Leila Amgoud (2003). A dialogue-game protocol for agent purchase negotiations. *Journal of Autonomous Agents and Multi-Agent Systems* 7(3): 235–273.

Meijers, Anthonie (1999). *Belief, Cognition and the Will*, chapter Believing and Accepting as a Group, pp. 59–71. Tilburg University Press.

Meijers, Anthonie (2002). Collective agents and cognitive attitudes. *Protosociology* 16: 20–85.

Meijers, Anthonie (2003a). Can collective intentionality be individualized? *American Journal of Economics and Sociology* 62: 167–183.

Meijers, Anthonie (2003b). Why accept collective beliefs? reply to gilbert. *Protosociology* 18–19: 377–388.

Meyer, John-Jules Ch. (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.

Meyer, John-Jules Ch. and Wiebe van der Hoek (1995). *Epistemic logic for AI and computer science*. Cambridge University Press, Cambridge.

Millikan, Ruth G. (1984). *Language, thought, and other biological categories*. MIT Press, Cambridge.

Millikan, Ruth G. (1990). *White queen psychology and other essays for Alice*. Cambridge. MIT Press.

Molden, Daniel C. and E. Tory Higgins (2005). *The Cambridge handbook of thinking and reasoning*, chapter Motivated thinking, pp. 295–317. Cambridge University Press, Cambridge.

Neches, Robert, Richard Fikes, Tim Finin, Thomas Gruber, Ramesh Patil, Ted Senator, and Williams R. Swartout (1991). Enabling technology for knowledge sharing. *AI Magazine* 12(3): 36–56.

Nickles, Matthias (2005). Exposing the communication level of open systems: Expectations, ostensible attitudes and multi-source assertions. Technical report, University of Munich.

Nickles, Matthias, Felix Fischer, and Gerhard Weiss (2005). A Framework for the Representation of Opinions and Ostensible Intentions. In van der Hoek, Wiebe, Alessio Lomuscio, Erik Vink, and Mike Wooldridge, editors, *Int. Workshop on Logic and Communication in Multiagent Systems (LCMAS 2005)*. ENTCS.

Noriega, Pablo and Carles Sierra (2002). Electronic institutions: Future trends and challenges. In *LNAI 2446*, Berlin. Springer-Verlag.

North, Douglass C. (1990). *Institutions, Institutional Change, and Economic Performance*. Cambridge University Press, Cambridge.

Object Management Group, OMG (2003). Object constraint language specification 1.4. `http://www.omg.org/`.

Ortony, Andrew, Gerald L. Clore, and Allan Collins (1988). *The cognitive structure of emotions*. Cambridge University Press.

Pasquier, Philippe (2005). Aspects cognitifs des dialogues entre agents artificiels : l'approche par la cohérence cognitive. Ph.D. diss., Faculté des sciences et de Génie, Département d'informatique et de génie logiciel, Université Laval, Quebec, Canada.

Pasquier, Philippe, Mathieu Bergeron, and Brahim Chaib-draa (2004). Diagal : a generic acl for open systems. In Gleizes, M.-P., A. Omicini, and F. Zambonelli, editors, *Proc. Fifth International Workshop Engineering Societies in the Agents World (ESAW'04)*, Vol. 3451 of *LNAI*, pp. 139 – 152. Springer-Verlag.

Pasquier, Philippe, Roberto Flores, and Brahim Chaib-draa (2004). Modelling flexible commitments and their enforcement. In Gleizes, M.-P., A. Omicini, and F. Zambonelli, editors, *Proc. Fifth International Workshop Engineering Societies in the Agents World (ESAW'04)*, Vol. 3451 of *LNAI*, pp. 153 – 165. Springer-Verlag.

Paurobally, Shamimabi, Jim Cunningham, and Nicholas R. Jennings (2005). A formal framework for agent interaction semantics. In *AAMAS '05: Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*, pp. 91–98, New York, NY, USA. ACM Press.

Pettit, Philip (2001). Deliberative democracy and the discursive dilemma. *Philosophical Issues* 11: 268–99.

Platon (1999). *Menon*. Le Livre de Poche.

Polinsky, A. Mitchell and Stephen Shavel (1998). *The New Palgrave Dictionary of Economics and The Law*, chapter Punitive Damages, pp. 192–198. Number 3. Macmillan Reference Limited, London.

Posner, Richard A. and Eric B. Rasmusen (1999). Creating and enforcing norms, with special reference to sanctions. *International Review of Law and Economics* 19(3): 369–382.

Quinton, Anthony (1976). Social objects. *Proceedings of the Aristotelian Society* LXXVI: 1–27.

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

Reed, Chris (1998). Dialogue frames in agent communication. In *Proceedings of the 3rd International Conference on Multi Agent Systems (ICMAS98)*, pp. 246–253, Paris , France. IEEE Computer Society Press.

Royakkers, Lambèr and Franck Dignum (2000). Organizations and collective obligations. In Ibrahim, M., J. Küng, and N. Revell, editors, *Proceedings of DEXA 2000*, Vol. 1873 of *LNCS*, pp. 302–311. Springer-Verlag.

Russell, Bertrand (1989). *Écrits de logique philosophique*. PUF. (Traduction française).

Sadek, David (1991). Attitudes mentales et interaction rationnelle : vers une théorie formelle de la communication. Ph.D. diss., Université de Rennes I, Rennes, France.

Sadek, David (1992). A study in the logic of intention. In Nebel, Bernhard, Charles Rich, and William Swartout, editors, *Proc. Third Int. Conf. on Principles of Knowledge Representation and Reasoning (KR'92)*, pp. 462–473. Morgan Kaufmann Publishers.

Sadek, David (2000). Dialogue acts are rational plans. In Taylor, M.M., F. Néel, and D.G. Bouwhuis, editors, *The structure of mutimodal dialogue*, pp. 167–188, Philadelphia/Amsterdam. John Benjamins Publishing Company.

Sahlqvist, Henrik (1975). Completeness and correspondence in the first and second order semantics for modal logics. In Kanger, S., editor, *Proc. 3rd Scandinavian Logic Symposium*, Vol. 82 of *Studies in Logic*.

Scheff, Thomas (1967). Toward a sociological model consensus. *American Sociological Review* 32: 32–46.

Schelling, Thomas C. (1960). *The strategy of conflict*. Harvard University Press, Cambridge.

Schiffer, Stephen (1972). *Meaning*. Oxford University Press, Oxford.

Schwitzgebel, Eric (2002). A phenomenal, dispositional account of belief. *Nous* 36: 249–275.

Schwitzgebel, Eric (2006). Belief. Stanford Encyclopedia of Philosophy.

Searle, John R. (1969). *Speech acts: An essay in the philosophy of language*. Cambridge University Press, New York.

Searle, John R. (1979). *Expression and Meaning. Studies on the Theory of Speech Acts*. Cambridge University Press.

Searle, John R. (1983). *Intentionality: An essay in the philosophy of mind*. Cambridge University Press.

Searle, John R. (1995). *The Construction of Social Reality*. Free Press, New York.

Searle, John R. and Daniel Vanderveken (1985). *Foundation of illocutionary logic*. Cambridge University Press.

Singh, Munindar P. (1991). Social and psychological commitments in multiagent systems. Technical report TM-91-08, Germany.

Singh, Munindar P. (1998). Agent communication languages: Rethinking the principles. *Computer* 31(12): 40–47.

Singh, Munindar P. (2000). A Social Semantics for Agent Communication Languages. In Dignum, Frank and Mark Greaves, editors, *Issues in Agent Communication*, number 1916 in LNAI, pp. 31–45. Springer-Verlag.

Stalnaker, Robert C. (1984). *Inquiry*. MIT Press, Cambridge.

Tollefsen, Deborah Perron (2002). Challenging epistemic individualism. *Protosociology* 16: 86–117.

Tollefsen, Deborah Perron (2003). Rejecting rejectionism. *Protosociology* 18–19: 389–405.

Traum, David R. (1994). Computational theory of grounding in natural language conversation. Ph.D. diss., Computer Science Departement, University of Rochester.

Traum, David R. (1999). Speech acts for dialogue agents. In Wooldridge, M. and A. Rao, editors, *Foundations of Rational Agency*, pp. 169–201. Kluwer Academic Publishers.

Tuomela, Raimo (1984). *A Theory of Social Action*. Reidel Publishing Company, Dordrecht and Boston.

Tuomela, Raimo (1992). Group beliefs. *Synthese* 91(3): 285–318.

Tuomela, Raimo (1995). *The Importance of Us: A Philosophical Study of Basic Social Notions*. Stanford Series in Philosophy. Stanford University Press.

Tuomela, Raimo (2000). Belief versus acceptance. *Philosophical Explorations* 2: 122–137.

Tuomela, Raimo (2002). *The Philosophy of Social Practices: A Collective Acceptance View*. Cambridge University Press, Cambridge.

Tuomela, Raimo (2004). *International Encyclopedia of the Social & Behavioral Sciences*, chapter Shared Belief, pp. 14039–14043. Elsevier.

Tuomela, Raimo (2005). We-intentions revisited. *Philosophical Studies* 125(3): 327–369.

Tuomela, Raimo (2007). *The Philosophy of Sociality*. Oxford University Press, Oxford.

Ullmann-Margalit, Edna and Margalit Avishai (1992). Holding true and holding as true. *Synthese* 92(2): 167–187.

van Ditmarsch, H.P., W. van der Hoek, and B.P. Kooi (2005). Public announcements and belief expansion. In Schmidt, R., I. Pratt-Hartmann, M. Reynolds, and H. Wansing, editors, *Advances in Modal Logic*, Vol. 5, pp. 335–346, London. King's College Publications.

Van Frassen, Bas C. (1980). *The Scientific Image*. Clarendon Press, Oxford.

Vanderveken, Daniel (1990). *Principles of language use*, Vol. 1 of *Meaning and Speech Acts*. Cambridge University Press.

Vanderveken, Daniel (1991). *Formal semantics of success and satisfaction*, Vol. 2 of *Meaning and Speech Acts*. Cambridge University Press.

Velleman, J. David (2000). *The Possibility of Practical Reason*. Oxford University Press, Oxford.

Verdicchio, Mario and Marco Colombetti (2003). A Logical Model of Social Commitment for Agent Communication. In *Proc. Second Int. Joint Conf. on Autonomous Agents and MultiAgent Systems (AAMAS-2003)*, pp. 528–535. ACM Press.

Vold, Georges B., Thomas J. Bernard, and Jeffrey B. Snipes (2002). *Theoretical Criminology*. Oxford University Press, fth edition.

Walton, Douglas N. and Erik C. W. Krabbe (1995). *Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning*. State University of New-York Press, NY.

Williams, Bernard (1970). *Problems of the self*, chapter Deciding to believe, pp. 136–151. Cambridge University Press.

Wooldridge, Michael (2000). *Reasoning about Rational Agents*. MIT Press.

Wray, K. Brad (2001). Collective belief and acceptance. *Synthese* 3(129): 319–333.

Wray, K. Brad (2003). What really divides gilbert and the rejectionnists. *Protosociology* 18–19: 363–376.

Zaibert, Leo A. (2003). Collective intentions and collective intentionality. *American Journal of Economics and Sociology* 62(1): 209–232.

TITRE :
Formalizing social attitudes in modal logic

ABSTRACT :
One of the most powerful tools to explain and predict an agent's behavior is to describe him thanks to his mental states, such as his beliefs or his intentions. In Artificial Intelligence, many researchers have focused on the formalization in modal logic of these individual mental attitudes, in order to use them in artificial agents. Lots of examples, such as: « The government believes that war will begin soon. », highlight the fact that attitudes, and beliefs in particular, can be ascribed to a group of agents. Besides it is interesting to notice that, even if the government as a whole believes that war will begin soon, some government members can disagree privately. The first aim of this dissertation is to provide a logical framework to represent the concept of group belief and to describe its features and its links with individual mental attitudes. It also appears that group belief in this sense results from a debate between group members. The second aim of this dissertation is thus to highlight the close link existing between group belief, dialog and speech acts.

AUTEUR : Benoit Gaudou

TITRE : Formalisation en logique modale d'attitudes sociales

DIRECTEURS DE THESE : Andreas Herzig et Dominique Longin

LIEU ET DATE DE SOUTENANCE : 10 juillet 2008 à l'IRIT

RESUME :

Décrire un agent à l'aide de ses états mentaux, comme ses croyances ou ses intentions, est un des moyens les plus puissants pour expliquer ou prédire son comportement. En intelligence artificielle, de nombreuses recherches ont été mené pour décrire en logique (notamment en logique modale) ces attitudes mentales individuelles dans le but de les intégrer dans des agents artificiels. De nombreux exemples, comme : « Le gouvernement pense qu'une guerre est inévitable », illustrent le fait que des attitudes mentales, en particulier des croyances, peuvent être attribuées à un groupe d'agent. Il est intéressant de constater que même si le groupe appelé « gouvernement », formé de ministres, croit que la guerre est inévitable, certains ministres peuvent avoir un avis privé différent. Le but de cette thèse est donc de donner une représentation en logique modale de la croyance de groupe, d'en décrire les propriétés logiques et les liens qu'elle entretient avec les attitudes individuelles. Il apparait que la croyance de groupe dans ce sens émane souvent d'une discussion entre les différents membres du groupe pour arriver à un compromis. Nous montrerons donc les liens étroits qu'elle entretient avec le dialogue et les actes de langages.

MOTS-CLES :

Croyance de groupe, logique modale, actes de langage, Langages de Communications entre Agents, Systèmes Multi-Agents

---

DISCIPLINE ADMINISTRATIVE :

Informatique – Intelligence Artificielle

ADRESSE DU LABORATOIRE :

Institut de Recherche en Informatique de Toulouse,

Université Paul Sabatier, 118 Route de Narbonne, F-31062 TOULOUSE CEDEX 9
