TITLE: A LOGICAL APPROACH TO FORMALISING NEGOTIATION IN MULTI-AGENT SYSTEMS

Running Head: Negotiation in Multi-Agent Systems

Authors and Affiliations:

Pietro Panzarasa
Department of Electronics and Computer Science
University of Southampton
Southampton SO17 1BJ, U.K.
pp@ecs.soton.ac.uk

Kathleen M. Carley
Department of Social and Decision Sciences
Carnegie Mellon University
Pittsburgh, PA 15213 USA
kathleen.carley@cmu.edu

Nicholas R. Jennings
Department of Electronics and Computer Science
University of Southampton
Southampton SO17 1BJ, U.K.
nrj@ecs.soton.ac.uk

Address for correspondence:

Prof. Nicholas R. Jennings
Department of Electronics and Computer Science
University of Southampton
Highfield, Southampton SO17 1 BJ, U.K.
nrj@ecs.soton.ac.uk
Tel: +44 23 8059 7681
Fax: +44 23 8059 3313

ABSTRACT

A new view of multi-agent systems is emerging within a variety of scholarly fields, from distributed artificial intelligence to sociology and organisation science. The hallmark of this view is a recognition of multi-agent systems as inherently complex and computational systems in which cognition occurs at multiple levels, not only within the individual agent, but as an emergent phenomenon from the interaction among multiple agents. According to this perspective, elucidating mechanisms of joint behaviour that are predicated on fully explicated models of agents' cognition at the individual and joint levels remains a key problem for multi-agent system research and practice. In this paper, a step in this direction is taken. We present a theory of negotiation, herein conceptualised as a mechanism for the social and cognitive integration of socially and cognitively differentiated agents. To this end, a new quantified multi-modal logical language is developed that allows us to reason about and represent agents' mental attitudes and their interplay. Drawing on this language, a model of negotiation is then formalised using the classical axiomatic-deductive methodology for theory building. Assumptions of the model are presented, and properties are discussed on a proof-theoretic basis. The descriptive breadth of the theory is illustrated by looking at its computational benefits and applicability when faced with a variety of real-world constraints, such as the agents' bounded rationality, informational asymmetry, and opportunistic behaviour.

KEYWORDS: automated negotiation; cognitive and social agenthood; artificial agents; BDI logic; joint commitment; cognitive and social integration

1. Introduction

Theories of negotiation abound, ranging from those derived from mainstream computer science and distributed artificial intelligence (DAI) (e.g. [8], [25], [26], [41], [51], [68], [74], [75]), to those that use microeconomic and/or statistical methods (e.g. [7], [46]), to those arising from more traditional social and organisational perspectives, such as organisation theory (e.g. [44], [47]), and sociology (e.g. [4], [29], [66]). Despite attempts to account for the individual agent's information-processing capabilities, knowledge and social abilities, none of these views is based on a fully explicated and plausible model of *cognitive and social agenthood*. In most of these studies, cognition and social behaviour appear to be two dichotomic conceptual constructs that can reluctantly be addressed under a unified theoretical perspective. On the one hand, when the focus is primarily on the organisation and management of the social relations between two or more negotiating agents, researchers tend not to embrace a cognitively oriented conceptual framework, and therefore concepts such as inter-agent social behaviour are left ill-defined with respect to the agent's mental state. On the other, when the emphasis is on the negotiating agent's mental state and cognitive architecture, the role of the agent's social ability tends to be underestimated, and little attention is paid to the cognitive representation of the social environment in the agent's mind.

Recent advances in social networks, cognitive sciences, DAI, and organisation theory have led to a new perspective on multi-agent systems that takes into account both their computational nature and the underlying social and knowledge networks (e.g. [39], [43], [73]). At the heart of this perspective is the argument that cognition occurs at multiple levels, not only within the individual agent, but as an emergent phenomenon from the interaction among multiple agents. Along these lines, attempts have been made to extend models of individual cognition and behaviour to a social setting in which collective behaviour is modelled as grounded on mutual beliefs, and motivated by joint goals, joint intentions and joint commitments (e.g. [49], [73]). At this level, sociality is described in terms of a higher-order joint mental state and joint behavioural processes that emerge from and transcend the mental states and behaviours of cognitively and socially interconnected individual agents.

Drawing on this view of cognitive and social agenthood, in this paper we develop a theory of negotiation in a multi-agent setting, where mechanisms of joint behaviour are predicated on a conceptually explicated model of the agents' cognition at both the individual and the joint level. To this end, the negotiating agent is conceptualised not only as an intentional system, but also as a theoretically richer cognitive entity that is able to reciprocate its intentional stance and model its acquaintances as also being intentional systems [24]. Building on this conceptual framework, we will show how it is possible to take some steps towards a theory of the *cognitive foundations for pluralism* in social systems. This is the main original theoretical contribution of the paper, besides its aim of presenting a new paradigm for conceptualising the negotiation behaviour within a multi-agent environment.

On more methodological grounds, the paper provides a new answer to the problem concerning what language, tools, and methods to use for theorising about social behaviour. We present and apply a computational logic for multi-agent systems that is intended to be a refinement of a standard Belief-Desire-Intention (BDI) logic. More specifically, we develop a first-order, linear-time, many-sorted, multi modal logic that is enriched by a number of modal operators for representing and reasoning about cognition and behaviour both at the individual and the social level. Drawing on this computational Multi-Agent System (MAS) logic, we will formalise a model of negotiation using the classical axiomatic-deductive methodology for theory building [55]. Assumptions of the model will be presented, and properties discussed on a proof-theoretic basis. In this respect, one major methodological contribution of the paper is to show how using a computational MAS logic and modelling techniques for theorising about negotiation can enhance the consistency, clarity and soundness of the argumentation [53]. In addition, making the underlying argumentation structure of the theory more explicit will, in turn, shed light on its descriptive and prescriptive power.

Finally, in endorsing a computational perspective for social theorising purposes, we make a further methodological contribution. Very much in the spirit advocated by many of the leading scholars in computational organisation science (e.g. [9], [14], [15]), we would like to take some steps towards an interdisciplinary integration of methods, analytical tools and research questions from differing theoretical perspectives such as computer science, DAI, organisation theory and sociology. By building on their connections, in this paper an attempt will be made to integrate these research areas towards a unified theoretical paradigm for the conceptualisation of pluralistic forms of organising social action. In bringing the core topics and theoretical constructs of some research lines (e.g. bounded rationality, opportunistic behaviour, social influence) to bear on the methodologies provided by others (e.g. computational logic, theorem-proving), we propose a new computational socio-knowledge logic-based perspective for describing, reasoning about, specifying, and ultimately verifying the properties of negotiation among multiple intelligent agents. Besides posing stimulating challenges for future research on negotiation, this new perspective provides us with a fundamental methodological question: whether and to what extent there should be as many paradigms in the study of pluralistic social systems as there are different scientific disciplines. To a certain extent, even the organisation of the scientific community seems to be at issue. In fact, working towards a unified paradigm is sympathetic with the idea that MAS theory and practice, rather than a body of research within DAI [37], represents an overarching perspective that integrates, under a common object of study, principles and methods from different research areas.

The remainder of the paper is organised as follows. The next section provides the background to our theory and introduces most of the ideas that will be developed in the paper. Section 3 presents an overview of the BDI logical framework: syntax, semantics, and basic properties of the logic will be briefly discussed. Section 4 is dedicated to a model of negotiation, which is basically organised as a set of assumptions

expressed in terms of our logic. In Section 5, properties of the model will be introduced and formalised as theorems that are logical consequences of the set of premises. Finally, Section 6 will summarise our findings and provide avenues for future work.

2. A Theory of Negotiation

This section provides some background to our formal theory of negotiation. The theory contains three main components. The first is a quantified multi-modal logical *language* that allows us to reason about and represent the mental attitudes and actions of agents (Section 3). The second is a model of negotiation which consists of a set of premises (empirical generalisations) formalised using the logical language (Section 4). Finally, the third component is the set of theorems that are logical consequences of the set of premises (Section 5).

At the outset, it is important to highlight what the theory developed in this paper is and is not intended to be. The theory is not primarily intended to model negotiation among human actors. The phenomena of human beliefs, intentions, preferences, commitments and actions are not the main concern of our study. Therefore, the theory should not be viewed as prescriptively conveying a logic with which a human agent should reason. Instead, our theory is mainly intended to conceptualise basic principles governing intelligent communities of artificial agents who negotiate in order to achieve a common objective. Such principles provide specifications for the design of artificial agents, and approximate a theory of human negotiation. As a result, to the extent that it provides assistance to practitioners interested in building distributed agent architectures, our theory is intended to contribute to DAI research. To the extent that it is flexible enough for describing and reasoning about the cognitive and behavioural properties of pluralistic systems comprising boundedly rational human agents, our analysis may contribute to organisation theory and cognitive science.

Even though this is not a theory of human negotiation, we have nevertheless drawn on the study of humans in developing this formal theory. Specifically, we draw on the research in cognitive science, organization theory (particularly computational organization theory), and social networks to provide guidance in our development of a new multi-agent logic that can be used for negotiation. Drawing on work in cognitive science and organization theory, we utilise the notion of bounded rationality [63]. Artificial agents will be limited in their ability to access and process information and so will need to generalise on the basis of empirical evidence; e.g., they might use machine learning techniques to estimate the state of another agent. Drawing on work in organization science, we utilise the notions of shared mental models, consensus formation and team decision making to characterize the ability of agents to make decisions as a group (the negotiation) without having a complete mapping of the other agents' beliefs, intentions, social commitments, and so forth. From the mental attitudes of the negotiating agents emerges a set of mutual beliefs, joint intentions, and joint commitments, i.e., a shared mental model, that is only partially shared and that is used in

the negotiation to reach (or not) consensus (agreement). Finally, based on constructural theory (e.g. [11], [12]) we take the stance that the agents' mental states and social position (behaviour) co-evolve, creating a corresponding co-evolution of the collective and agent levels. Social behaviour is mediated by cognition concurrently for all negotiating agents. Thus, it is the perception of others' interactions, attitudes, and goals that matters to the agent's behaviour. Finally, we note that simple negotiation is a fairly simple social task. As noted in [13], negotiation requires agents who are at least rational and engaged in a multi-agent situation. This makes negotiation a reasonable task to use for developing new multi-agent logics.

Negotiation is here conceived of as a *socio-cognitive mechanism for the transformation of commitments* in a social setting comprising cognitive agents [49]. By this, we mean that negotiation is grounded on a joint commitment among the members of a group to achieving a state of the world, and it is aimed at generating a joint commitment among the agents to acting according to a joint plan of action in order to achieve that state. By describing negotiation as triggered by a joint commitment, we can give a comprehensive and flexible account of a range of negotiation schemes, from those in which the agents represent opposing forces and have contrasting objectives and views, and competitive interests, through to those in which the agents have consistent objectives and views, and overlapping interests. To illustrate this point, consider the case of an economic transaction, in which the agents usually undertake two possible roles: buyers and sellers. In general, buyers and sellers have opposing views and goals, e.g. the buyer wants a low price, whereas the seller tries to obtain the highest price. However, both the buyer and the seller are moved into negotiation by a joint commitment towards a common state of the world – that in which the economic transaction has taken place, i.e. a state in which the good (or service) has been delivered from the seller to the buyer, and its counter-value from the buyer to the seller. If successful, negotiation will integrate the agents' opposing goals and views as to how to fulfil their joint commitment. A new joint commitment will eventually ensue towards a new common state of the world – that in which both the buyer and the seller intend to carry out a joint plan aimed at finalising the economic transaction.

This way of describing negotiation in terms of transformation of commitments is sympathetic with the idea that negotiation is inherently intertwined with a problem-solving activity undertaken in a social setting [49]. In this view, negotiation can be seen as a *mechanism for finding a solution to a common problem in a collaborative manner*. The common problem concerns what is to be done by the group in order to attain a state of the world; the solution reflects an agreement among the group as to the most appropriate joint plan of action that achieves that state (e.g. [49], [75]). This way of describing negotiation allows us to highlight how intimately the process rests on that form of reasoning that is usually intended to be directed towards conduct and particularly towards the resolution of a practical problem, namely practical reasoning [3]. In essence, practical reasoning helps coordinate intentions towards a state of affairs with possible alternative courses of action that are means to achieve that state. Therefore, practical reasoning is an

essential ingredient of negotiation. It works out the steps towards the final agreement by forming the sequences of possible paths of actions that each agent deems appropriate to a situation.

If negotiation is successful, its outcome is an agreement on a joint plan that represents an acceptable means for fulfilling the joint commitment to achieving a common state. Such an agreement is a composite concept in that it reflects what practical judgements (i.e., the belief about what is to be done) the agents have brought about via practical reasoning and to what extent the agents have compromised in order to find an overlapping area between their differing judgements. An agreement also reflects a joint commitment to acting in accordance with the agreed-upon plan. Coming to an agreement about a plan thus transforms a joint commitment to finding a plan for achieving a state into a joint commitment to performing a plan. This transformation of commitments is what constitutes the essence of negotiation.

3. The Language

This section gives an overview of the formal framework in which the model of negotiation will be expressed. The logic is a simplified version of what we developed in [49] to which the reader should refer for a complete formal definition. This framework is a first-order, liner-time, quantified, many-sorted, multi-modal logic for reasoning about agents, groups, actions, and mental attitudes, with explicit reference to time points and intervals. The logic is first-order, in that it allows quantification over terms. It is multi-modal, in that it is enriched by some further modal connectives for referring to the beliefs, intentions, commitments and preferences of agents. Furthermore, the logic contains a simple apparatus for representing the actions performed by agents, which makes use of ideas from dynamic logic [31].

First, a brief description of the model of time that underpins our logic. Every occurrence of a formula ϕ is stamped with a time *t*, written $\phi(t)$, meaning that ϕ holds at time *t*. Time is taken to be composed of points and, for simplicity, is assumed to be discrete and linear. Temporal intervals are defined as pairs of points. Intervals of the form *(t, t)* can equally be written as time points[1]. The usual connectives of linear temporal logic - $\Diamond \phi\ (t)$ (meaning ϕ is eventually satisfied); $\_\phi\ (t)$ (meaning ϕ is always satisfied); $U(\phi, \psi)\ (t)$ (meaning ψ is satisfied until ϕ becomes satisfied); $W(\phi, \psi)\ (t)$ (meaning ψ is satisfied unless ϕ becomes satisfied) - can be defined in the following way (e.g. [5], [16], [35]):

$\Diamond \phi\ (t) \equiv \exists\ t'$ s.t. $(t < t' \wedge \phi\ (t'))$

$\_\phi\ (t) \equiv \forall\ t'\ (t < t' \wedge \phi\ (t'))$

$U(\phi, \psi)\ (t) \equiv \exists\ t'$ s.t. $(t < t' \wedge \phi\ (t') \wedge \forall\ t''\ (t \leq t'' < t' \supset \psi\ (t'')))$

$W(\phi, \psi)\ (t) \equiv U(\phi, \psi)\ (t) \vee \_\ \psi\ (t)$

[1] In what follows we will adopt the convention that a missing temporal term is the same as the closest temporal term to its right. For

The logic is many-sorted. Terms come in six sorts:

- First, we have terms that denote *agents*, and we use a_i, a_j, … and so on as variables ranging over individual agents.
- Second, we have terms that denote *groups* of agents, and we use *gr*, *gr'*, … and so on as variables ranging over such groups. A group *gr* of agents is simply a non-empty subset of the set of agents. Agents and groups may easily be related to one another via simple set theory. With the operator, we relate agents to groups of agents: $a \in gr$ means that the agent denoted by *a* is a member of the group denoted by *gr*.
- Third, we have terms that denote *sequences of actions*, and we use e_h, e_k, … and so on as variables ranging over sequences of actions. Action sequences may be distinguished depending on whether they can be performed by an individual agent (single-agent actions) or by a group of agents (multi-agent actions). To simplify the specification, we assume that an action sequence is either single-agent or multi-agent, but not both. In addition, we introduce the notion of *plan*, as a state-directed action that an agent or a group of agents can perform in order to achieve a state of the world. The operator *plan(gr, e, ϕ) (t)* represents a multi-agent action sequence which a group *gr* can perform at time *t* in order to bring about a state ϕ [2]. As happens with the broader category of actions, plans may be either single-agent or multi-agent, but not both. Further, we distinguish between a plan abstraction - *plan(gr, e, ϕ) (t)* - and its occurrence in the world - *<plan(gr, e, ϕ)> (t)*. The former refers to an action sequence that *might* be invoked by agents or groups to satisfy certain states of the world. The latter refers to the execution of a plan by agents or groups, which in turn results in the occurrence of an action sequence and the subsequent achievement of a state of affairs as a consequence of the performance of that action sequence.
- Fourth, we have terms that denote *time points*, and we use *t*, *t'*, … and so on as variables ranging over time points.
- Fifth, we have terms that denote *temporal intervals*, and we use *i*, *i'*, … and so on as variables ranging over time intervals. Temporal intervals are defined as pairs of time points: *(t, t')*, *(t', t'')*, …
- Finally, we have terms that denote other generic objects in the environment, and we use o_a, o_b, … and so on as variables ranging over objects.

---

example, *Bel* (a_1, *Int* (a_2, ϕ)) *(t)* states that at time *t* agent a_1 believes that at time *t* agent a_2 intends to attain ϕ.

[2] A more sophisticated definition of plans could have been adopted. For example, it might be useful to distinguish between the body and the preconditions of a plan. Moreover, we could have made explicit representations of partial plans as well as hierarchical non-linear plans (e.g. [28], [38]). However, such refinements are not relevant for our current analytical purposes, and we leave them to future work.

As discussed in [49], the logic is enriched by a set of modal operators for reasoning about agents' mental attitudes. First, we have the operators *Bel(a, ϕ) (t)* and *Int (a, ϕ) (t)*, which mean that at time *t* agent *a* has, respectively, a belief that ϕ holds and an intention towards ϕ, where ϕ is an atomic proposition[3]. An agent's belief set includes beliefs concerning the world, beliefs concerning mental attitudes of other agents, and introspective beliefs (i.e., beliefs can be nested). This belief set may be incomplete. An agent may update its beliefs by observing the world and by receiving messages from other agents. The formal semantics for beliefs are a natural extension of the traditional Hintikka possible-worlds semantics (e.g. [33], [34]). The restrictions to be imposed on the belief-accessibility relation ensure a belief axiomatisation of KD45 (corresponding to a "Weak S5 modal logic"), which thus implies that beliefs are consistent and closed under implication, and that an agent is aware of what it does and does not believe [30].

Turning to intentions, the idea is that an agent's intention represents those states of the world that the agent is "self-committed" to achieving [77]. That is, as long as an agent intends to achieve a state, it has committed itself to perform all those actions that it deems appropriate for achieving that state. As with beliefs, intentions can be nested, i.e., their argument can be another modal operator (e.g. *Int (a_i, Bel (a_j, ϕ)) (t)*, meaning that at time *t* agent a_i intends that agent a_j believes that ϕ holds). Also, as with beliefs, the semantics of intentions are given in terms of possible worlds. Restrictions on the intention-accessibility relation ensure that the logic of intentions validates axioms K and D, which means that intentions are closed under implication, and are consistent. In addition, we introduce a strong realism constraint, which ensures that the agent's intentions do not contradict its beliefs [60].

In addition to beliefs and intentions, agents have local preferences. The operator *Pref(a, ϕ, ψ) (i)* means that agent *a* prefers ϕ over ψ at interval *i*, where ϕ and ψ are atomic propositions. Preferences can be nested and have other modal operators as their arguments (e.g. *Pref (a, plan(gr, e_h, ϕ), plan(gr, e_k, ϕ)) (t)*, meaning that agent *a* at time *t* prefers the joint plan containing action sequence e_h rather than the joint plan containing action sequence e_k). As we will see in Section 4, an agent's preference plays an active role in practical inferences, where a plan is to be selected in order for a given intention to be fulfilled. The semantics for preferences are given in terms of closest worlds (see [6] and [49] for details). Intuitively, agent *a* in current world *w* prefers ϕ to ψ at *i* - *Pref (a, ϕ, ψ)(i)* - iff the value $p \in \Re$ that *a* associates to the set of closest worlds to *w* in which ϕ is true and ψ is false is greater than the value $p' \in \Re$ that *a* associates to the set of closest worlds to *w* in which ψ is true and ϕ is false.

Finally, we have the deontic operators *Comm(a, gr, ϕ) (t)* and *J-COMM(gr, ϕ)(t)*, which mean, respectively, that agent *a* is committed to group *gr* to achieving ϕ, and group *gr* is jointly committed to achieving ϕ (e.g. [49], [73], [75]). Intuitively, a commitment between an agent and a group (or among the

[3] Our analysis is based on a fairly standard BDI framework as found, for example, in [59] and [65]. For our purposes, we will restrict our consideration to the agent's beliefs and intentions. In addition, we will use a deontic operator to refer to the agents' commitments. Finally, we will introduce an operator to express the agent's preference over a pair of formulas.

members of a group) reflects an agreement between the agent and the group (or among the group members), and the right of the group to control the behaviour of the agent (or of its members) [49]. In addition, it reflects the obligation of the agent (or of the members) towards the group. These operators are derived from the primitive operator *Comm (gr, gr', e) (t)*, which expresses the relation between two groups of agents in terms of one group, *gr*, being committed to another, *gr'*, to performing action *e*. Full details of the semantics for these deontic operators are given in [49].

In addition to these modal operators, we have first-order equality: a formula ($\tau = \tau'$) will be true if τ and τ' denote the same individual. The operators $\neg$ (not) and (or) have classical semantics, as does the universal quantifier . The remaining classical connectives and existential quantifier are assumed to be introduced as abbreviations, in the obvious way. We also use the punctuation symbols ")", "(", "[", "]", and comma ",".

4. The Model

In this section we will outline our model of negotiation which will be expressed in the language introduced in Section 3. This model can be used for theoretical as well as practical purposes. First, from a theoretical perspective, properties are derived and proved from the assumptions of the model. This allows us to place the study of the ongoing social and cognitive processes that underpin negotiation on a more secure and formal footing. Second, the model can be used to guide the development of an agent architecture for practical purposes, and to undertake and evaluate a set of virtual experiments based on this architecture (e.g. [36], [40]). Here our objective is to develop a formalisation of negotiation that is comprehensive enough to account for a number of instantiations that might occur in real-world domains. Therefore, we will not express our model in terms of a set of computational tactics and strategies that the negotiating agents may use during their interaction [25]. Nor will we specify the logical modelling of the protocol that defines the various states in which an agent may be during a negotiation and thus the transition between states in which an agent may be involved [41]. Rather, we will formalise our model in terms of the weakest conditions under which a negotiation process can be said to have occurred (e.g. [49], [75]). This approach is motivated by a two-fold observation. First, it is broad enough to guarantee a comprehensive account of a wide range of negotiation processes occurring in differing domains and involving differing agents who employ differing behavioural strategies and computational techniques. Second, it is also specific enough to provide an understanding of the key underpinning structures and processes that appear to be common to most forms of negotiation.

In most real-world situations, negotiation is viewed as the process by which an agreement is made by two or more parties (e.g. [25], [36]). Agents usually make proposals and counter-proposals; they might suggest modifications and receive requests for amendments of their own proposals; they might have

objections to one or more of the alternative plans. Negotiation can range over a number of quantitative and qualitative aspects of plans (e.g. time; duration; price) Each successful negotiation is therefore expected to resolve a number of different issues to the satisfaction of each agent. Most of the time, a trade-off between contrasting issues might be required in order for the agents to come to an agreement [25].

In order for a group of agents to start negotiating with one another, they need to be jointly committed to attaining a state of the world. Negotiation will lead them to agree about which plan to perform in order to achieve that state, and to commit themselves to act in accordance with such a plan. Joint commitment is a composite concept that reflects a number of properties. Among these (see [49] for details), a joint commitment towards ϕ evokes a shared mental state in which each member of the group believes that: (i) ϕ is currently false; and (ii) ϕ will eventually be true. Both properties suggest that the agents' efforts to achieve state ϕ are not meaningless. In particular, (i) means that the agents are committed to achieving something that does not already exist. On the other hand, (ii) means that the agents' commitments are directed towards something that is *in theory* feasible. However, being feasible in theory does not mean being feasible *in practice*. That is, the agents might believe that a state will eventually be true, but might not have the ability to attain it themselves. More specifically, the members of a group jointly committed to attaining a state ϕ may believe that ϕ will be eventually true because: (a) either they believe that at least one of them holds a belief as to how the group can achieve ϕ; or (b) they believe that there is one or more agents outside the group that have the required ability and whose assistance can be asked for.

Joint commitment to achieving something must reflect either condition (a) or (b). Informally, this means that, as long as a group has committed itself to achieve a state, it has committed itself to find the means to bring about that state, and each member's believing that there are no such means, either inside or outside the group, would contradict their joint commitment. Now, since negotiation reflects joint commitment, the first minimum condition required for us to be able to say that negotiation occurred at all is that either condition (a) or condition (b) was true. However, as might be expected, condition (b) is *unstable* as it triggers group dynamics involving changes in the members until a new group is generated that has the necessary ability. Iteratively, the group will reconstruct itself until condition (a) becomes true. At this stage, the agents may start negotiating with one another[4]. Thus, against this background, the first stage of negotiation may be formally formulated. This stage reflects the changes in the group which lead to a state in which at least one member of the group believes that the group has the required ability:

[4] Note that once a joint commitment towards ϕ has been generated, cirmumstances may change. In Proposition 1, we express the fact that, although such a commitment was grounded on the belief that ϕ was attainable, the agents may well change their beliefs and come to realise that ϕ is no longer attainable. In such a case, they will drop their commitment. However, all the other properties incorporated in the notion of joint commitment must be dynamically maintained in order for negotiation to start. In particular, the agents must: (i) believe that they are committed to achieving a state ϕ that is currently false; (ii) intend to achieve ϕ; (iii) be socially committed to the group to fulfilling their intentions; (iv) believe that each member is socially committed; (v) and believe that (ii) will continue to hold until some escape conditions become true [49].

Proposition 1. Given a group of agents jointly committed to achieving a state of the world ϕ, a state will follow in which the agents drop their commitments and negotiation fails unless, possibly as a consequence of changes in the group size, at least one member holds a belief concerning the identity of a plan that the group might perform in order to attain ϕ :

$\_\ \forall\ gr, \forall\ t\ J\text{-}COMM\ (gr, \phi)\ (t) \supset \exists\ t' > t$ s.t.
$W([\exists\ gr' \supseteq gr, \exists a_i \in gr', \exists\ e_h$ s.t. $Bel(a_i, plan(gr', e_h, \phi))], \neg J\text{-}COMM\ (gr, \phi))\ (t')$

Proposition 1 means that the agents do not keep their joint commitment for ever. That is, agents are not taken to be fanatical, and will eventually drop their commitment unless one of them comes up with a potential way of fulfilling it [19]. Along these lines, Proposition 1 can be informally restated in terms of what can be thought of as the starting point of a process of negotiation: Agents may start negotiating with one another when at least one of them comes to *fully represent a potential candidate solution of the practical problem (i.e. a plan) in its mind*[5]. Negotiation will take place thereafter, as the agent with a mental representation (i.e., a belief) of a plan will elaborate on it, discover alternative potential plans, form a practical judgement, and try to communicate and make its acquaintances aware about it. A proposal will then be generated and sent onto the other agents, and such a proposal will eventually be evaluated by the other agents, and a series of counter-proposals may then be generated until an agreement is eventually made. These steps of the negotiation will be dealt with in turn in what follows.

Given a state of the world to be jointly achieved by a group of agents, a very large part of each member's effort in any negotiation is devoted to discovering possible alternatives of plans the group may perform to achieve that state [63]. When successful, this selective search activity typically originates either

[5] We recognise that a solution to a practical problem may be an *emergent* outcome of negotiation. The emergence perspective needs to be qualified in two respects. First, emergence means that, at the beginning of negotiation, none of the agents needs to have a precise idea as to what a potential outcome might be. This means that it need not necessarily be the case that a potential solution is represented in the mind of some agents in order for negotiation to start and continue. According to the emergence perspective, agents need only know that *some* solution exists, and they negotiate in order to eventually find out exactly which solution can solve their problem. A second tenet of the emergence perspective is that the final outcome of negotiation stems from a path of *partial* solutions that the agents communicate with one another until a final solution is agreed upon. Partial solutions require *partial plans*, that is, plans that help the agents to get closer to the fulfilment of their commitment, by achieving something that is instrumental to the final end. For the analytical purposes of our paper, the emergent properties of negotiation are not an object of our study. Firstly, our Proposition 1 explicitly requires the mental representation of a plan in an agent's mental state. This requirement draws heavily on the way we formalise the notion of ability. Specifically, we define ability in terms of quantification *de re* [35] (see the formalisation of *Can(a, ϕ)* in [49]). That is, action e_h in Proposition 1 is quantified *de re* with respect to the *Bel* modality (e.g. [16], [35]). This means that the agent must be "aware of the identity" of at least one solution in order to believe that a problem is solvable. On the other hand, emergence would simply require a notion of ability without any rigid designator of exactly which solution will solve the problem. An agent may simply be aware that *some* solution exists and will start negotiation in order to find out the identity of such solution. Secondly, our Proposition 1 explicitly requires full plans to be communicated among the agents, that is, plans that might represent the final solution of negotiation. In turn, this requirement draws upon the way we formalise the notion of plan. In our framework, plans are introduced simply as actions that agents may perform to achieve some state of the world. On the other hand, the emergence property of partial solutions would require the notion of partial plan, that is plans that, if performed, contribute to the achievement of some state, without achieving it directly. The agents may thus perform partial plans in order to progressively find out how to fulfil their joint commitment. We leave both the development of a model of negotiation from an emergence viewpoint, and the

one or more alternative plans that may represent alternative solutions to the practical problem. Depending on the number of the possible solutions discovered, two different types of practical reasoning processes ensue. On the one hand, if the number of potential available solutions is one, the agent will generate a *practical necessity judgement* that, unless that plan is performed, the group cannot fulfil its joint commitment. On the other, if there are a number of alternative potential solutions and the agent is aware of such a range of possibilities, then it will have to express a preference and make a choice in order to form a *practical satisfactory judgement* (e.g. [3], [17], [63]). This judgement incorporates the plan that the agent believes would most satisfactorily[6] take the group closer to the solution of the practical problem (e.g. [49], [62]). In the light of these observations, we can now engage in the formalisation of the negotiating agent's *social practical inference* [69], that is, the structure of propositions that correspond to the reasoning process that a member of a jointly committed group undertakes in order to give an answer to a practical problem. Proposition 2 is intended to formalise the cognitive link between the individual agent's intention to achieve a state collaboratively and the related practical judgement concerning how to achieve that state:

Proposition 2. Given a member of a group (jointly committed to achieving ϕ) that holds a belief concerning the identity of a possible plan that the group might perform in order to attain ϕ, a state will follow in which that member: (i) either believes that the plan is the only possible one available to attain ϕ, or (ii) prefers that the group performs that plan among alternative possible plans, or (iii) prefers that the group performs another alternative plan:

$$\_\ \forall gr, \forall a_i \in gr, \forall e_h, \forall t$$
$$[\textit{J-COMM}\ (gr, \phi)\ (t) \wedge \textit{Bel}(a_i, \textit{plan}(gr, e_h, \phi))] \supset \exists\ t' \geq t \text{ s.t.}$$
$$\textit{Bel}(a_i, (\Diamond\phi \Leftrightarrow \Diamond<\textit{plan}(gr, e_h, \phi)>))(t') \vee \wedge_{k \neq h} \textit{Pref}\ (a_i, \textit{plan}(gr, e_h, \phi), \textit{plan}(gr, e_k, \phi)) \vee \exists e_k \text{ s.t. } \wedge_{k \neq h} \textit{Pref}\ (a_i, \textit{plan}(gr, e_k, \phi), \textit{plan}(gr, e_h, \phi))\ (t')$$

The antecedent of the above material implication contains two components: a conative and a doxastic mental attitude. The conative component represents the major motivational premise of the agent's social practical inference, based on the group's joint commitment to achieving a state of affairs. The doxastic component has its roots in the agent's search activity aimed at discovering potential solutions of the practical problem. The consequent of the implication contains the minor premise of a social practical inference. This is a doxastic premise that conveys a belief (practical judgement) as to how the joint commitment expressed

corresponding refinement of the logical language, for future work.

[6] Note that we deliberately do not specify any functional form of the preference function that the agent may use to select a plan. Therefore, the concept of *satisficing* can be variously defined depending on the agent's style and preference function (e.g. [62], [63]). The agent may be self-interested and choose the plan that is most beneficial for itself, while fulfilling the joint commitment of the group. Or, it may be other-oriented and choose the plan that is most beneficial to the fulfilment of the joint commitment of the group.

within the major motivational premise can be satisfied. More specifically, the typical role of this belief is to trigger the cognitive path that leads the agent from an intention to achieve a state of the world to an intention favouring the performance of a plan that brings about that state.

Against this background, we are now in a position to give a more formal definition of the typical structure of an agent's social practical reasoning (e.g. [49], [69]). This is given by an inference that contains, minimally, three sorts of constituents: (i) a major premise comprising a joint commitment to achieving a state of the world ϕ (hence, the agent's intention to achieve ϕ in a collaborative manner; see definition of joint commitment in [49]); (ii) a minor premise containing the agent's practical judgement as to the most appropriate means to achieve ϕ; and (iii) a conclusion containing the agent's intention that the group performs the plan suggested in (ii). Depending on whether the solution is unique or chosen among a set of alternative ones, a social practical inference can be seen as a deductive or non-deductive (either inductive or abductive) scheme of social practical reasoning (e.g. [49], [69]). In either case, it is a means-end argument intended to represent a transformation of intentions, from a prior one towards an end to a conclusive one towards the means to secure the end. In the light of this, while Proposition 2 was intended to formalise the cognitive link between the two first premises of a social practical inference, Proposition 3 completes the representation of the cognitive path that underpins practical reasoning, formalising the link between the two premises and the conclusive intention supporting the performance of the plan favoured by the practical judgement:

Proposition 3. Given a member of a group (jointly committed to achieving ϕ) that holds a practical judgement, a state will follow in which that member has successfully conducted a social practical inference. That is, a state will follow in which the practical judgement has led the agent to come up with the corresponding practical inferential conclusion, either in a deductive or non-deductive manner:

$$\_\ \forall\, gr, \forall\, a_i \in gr, \forall\, e_h, \forall\, t$$
$$[\textit{J-COMM}\,(gr, \phi)\,(t) \wedge \textit{Bel}(a_i, (\Diamond\phi \Leftrightarrow \Diamond\langle \textit{plan}(gr, e_h, \phi)\rangle))(t) \vee \wedge_{k \neq h} \textit{Pref}\,(a_i, \textit{plan}(gr, e_h, \phi), \textit{plan}(gr, e_k, \phi))\,(t)] \supset \exists\ t' \geq t \text{ s.t.}$$
$$\textit{Int}(a_i, \Diamond\langle \textit{plan}(gr, e_h, \phi)\rangle)\,(t')$$

In Proposition 3 the dynamic internal articulation of a social practical inference pattern has been formalised leading an agent from the generation of practical judgement (in the antecedent) through to the generation of a new intention (in the consequent). More specifically, Proposition 3 means that whenever an agent of a jointly committed group either believes that there is only one plan that the group can perform to fulfil its commitment or prefers a plan over a range of alternative ones, it will eventually form an intention that the group performs that plan.

So far we have glossed over the socio-cognitive foundations of negotiation. To this end, we have considered the problem of an individual agent's social practical reasoning behind negotiation. However, in order for a group to successfully undertake negotiation, the members' practical reasoning processes must be socially interconnected in such a way that an agreement can be reached. This involves a *coordination problem* that is captured and formalised in the following propositions of our model.

Once one of the agents has conducted a practical reasoning process and thus generated an intention that the group performs some plan, negotiation proceeds with that agent's intending to generate a state in which all its acquaintances know about that intention [75]. In most cases, this is achieved by the agent's generating a proposal and sending a message to all others specifying what plan it believes the group should perform in order to attain a state of the world. This plan is the first candidate for being moved up to an agreed-upon plan status, that is a plan on which the preferences of all the other agents may converge. Expressing this more formally:

Proposition 4. If a member of a group jointly committed to attaining a state of the world holds the intention that the group performs a particular plan, then it will also intend to bring about a state where every member holds a belief about its intention that the group performs that plan[7]:

$$\_\ \forall\, gr, \forall\, a_i, a_j \in gr, \forall\, e_h, \forall\, t$$
$$[J\text{-}COMM(gr, \phi)(t) \wedge Int(a_i, \Diamond\!<plan(gr, e_h, \phi)>)(t)] \supset \exists\, t' \geq t \text{ s.t.}$$
$$Int(a_i, Bel(a_j, Int(a_i, \Diamond\!<plan(gr, e_h, \phi)>))) (t')$$

In Proposition 4, an agent's intention to make the group perform a plan triggers a subsequent intention to impact upon the acquaintances' mental states. This transformation of intentions is ultimately aimed at making the members of the group generate a belief about a new fact: a proposal as to how the group should fulfil its joint commitment. The intention to let somebody know something can be regarded as an instantiation of a more general attitude: the intention to make somebody adopt a mental attitude [50]. This is a key construct that lies at the heart of most social processes and inter-agent social behaviours. In fact, it can be seen as the cognitive source of a variety of social influence processes that the agents exert in order to impact upon each other's mental states. The agent who is subjected to social influence will typically change its mental state and adopt new *socially motivated* mental attitudes (see also Section 5.4). These are attitudes that are inherently motivated by social behaviour and that rest on the agents' representing each other in intentional terms (e.g. [24], [50]). In this view, it is the mere social nature of the agent that affects, and thereby alters, its mental state from what it would otherwise have been, had the agent not engaged in any

[7] Note that here we assume that the agent believes that its inferential conclusion will still be true at the (future) time point when everybody in the group knows about this conclusion.

form of social behaviour. Along these lines, Proposition 5 is intended to formalise an agent's adoption of a new socially motivated belief. In turn, this is formalised as the typical outcome of the social influence that an agent exerts upon another as a result of the former's intention to let the latter know something:

Proposition 5. If a member of a group jointly committed to attaining a state of the world holds the intention that the group performs a particular plan, and also intends that the other members generate a belief about its intention that the group performs that plan, then a state will follow in which each member maintains such a belief[8]:

$\_\ \forall\ gr, \forall\ a_i, a_j \in gr, \forall\ e_h, \forall\ t$

$[J\text{-}COMM(gr, \phi)(t) \wedge Int(a_i, \Diamond<plan(gr, e_h, \phi)>)(t) \wedge Int(a_i, Bel\ (a_j, Int(a_i, \Diamond<plan(gr, e_h, \phi)>)))] \supset \exists\ t' \geq t$ s.t. $Bel\ (a_j, Int\ (a_i, \Diamond<plan(gr, e_h, \phi)>))\ (t')$

In Proposition 5, the expected cognitive effects of individual social behaviour have been formalised through a three-way nested modal operator: the intention about somebody's belief about somebody's intention. The consequent of the material implication axiomatises the fact that the expected cognitive results equal the actual ones. These are formalised with a two-way nested modal operator: the belief about somebody's intention. As it turns out, the agent's intention in the antecedent activates the mechanisms and structures that enable the social nature of agenthood to impact upon the group members' mental states. As will be shown in what follows, impacting upon somebody's mental state is a step towards the generation of higher-level mental attitudes (e.g. joint intentions, joint commitments) that rest on and, at the same time, transcend identical individual ones [49] (see also Sections 5.1.and 5.4).

Once the members of the group have come to know about one of their acquaintances' intentions to make the group act in a specified way, they will evaluate this new piece of information. Each agent will then act in differing ways depending on the extent to which the received proposal is consistent with its own beliefs, intentions, preferences, evaluation procedures, behaviour tactics, etc. However, on a more general level, we can identify three main alternating reactions of the members to the disclosure of an agent's intention: (i) they may agree with their acquaintance and accept the proposed solution; (ii) they may have a different belief as to how the group should act, and therefore may intend to let the others know about a modified solution or a new solution at all; or (iii) they may reject the proposed solution and withdraw from the process of negotiation[9]. Formally, we have:

[8] We assume that the social influence exerted by the agent upon its acquaintances is successful, therefore generating an impact upon the others' mental states. However, there are obvious exceptions to this assumption, most commonly related to the inability of the agent to communicate and interact with the other agents in an effective manner (e.g. [50], [73]).

[9] Rejection without withdrawing may also be possible. Also, a response may well contain just a *critique* to the initial proposal without any suggestion of a specific different plan to be performed [51]. However, formalisation of such forms of reply would

Proposition 6. If a member of a group jointly committed to attaining a state of the world believes that one of its acquaintances intends that the group performs a particular plan[10], then a state will follow in which the agent will withdraw from negotiation by dropping its commitment to the group unless it intends that the group: (i) either performs the proposed plan; or (ii) acts differently by performing another plan:

$\_ \forall gr, \forall a_i, a_j \in gr, \forall e_h, \forall t$

$J\text{-}COMM(gr, \phi)(t) \wedge Int(a_i, \Diamond\langle plan(gr, e_h, \phi)\rangle)(t) \wedge Bel(a_j, Int(a_i, \Diamond\langle plan(gr, e_h, \phi)\rangle))(t) \supset \exists t' > t$ s.t.

$W([Int(a_j, \Diamond\langle plan(gr, e_h, \phi)\rangle) \vee \exists e_k, k \neq h,$ s.t. $Int(a_j, \Diamond\langle plan(gr, e_k, \phi)\rangle)], \neg Comm(a_j, gr, \phi))(t')$

If just a single agent withdraws from the process by dropping its commitment to the group, negotiation fails and no agreement is reached. Another group may then form (for example, with the remaining agents or with new members), and another process of negotiation may start. An agent may withdraw from negotiation for a number of different reasons. For instance, the agent's behaviour may be subjected to time constraints: given this, if an agreement is not reached by a certain time point, an agent may exit the group without completing the negotiation [25]. Or, the agent may eventually come up with a new goal and/or intention that is inconsistent with the state of the world to be achieved collectively by the group. This will bring about a state in which the agent is no longer motivated to be part of the group, and to continue the negotiation. Or, the agent may well come to believe that the state of the world to be achieved by the group via negotiation is no longer attainable. Again, in such a situation, there is no motivation for the agent to carry on the negotiation and to keep its commitment to the group.

Assuming that, upon updating its beliefs with the new piece of information concerning a member's proposal, no agent drops its commitment to the group, negotiation will continue. As stated in Proposition 6, each agent may either hold a different view as to how the group should act or agree with the received proposal. In the former case, the agent will form the corresponding intention that the group performs the plan that it believes to be the most satisfactory solution for fulfilling the joint commitment. Again, this intention generation replicates the practical reasoning process that we described in Proposition 3, and can thus be formalised along the lines of a practical inference (either deductive or non-deductive). As in Proposition 3, we have the same major premise incorporating a joint commitment, and therefore the agent's intention to achieve a state of the world in a collaborative manner. However, in this case the minor premise contains a different practical judgement, and correspondingly the inferential conclusion reflects a different intention. As

require consideration of a fully explicated model of argumentation. For simplicity, we will not specify such a model here since our concern is to merely identify a set of conditions that underpin the success of a wide range of negotiation processes. However, elsewhere [50] by running a set of computer simulations we illustrated how giving additional information (e.g. explaining why a proposal or counter-proposal was made) can impact upon the outcome of the negotiation.

[10] Note that we do not assume that the agent believes correctly, namely that it is true that its acquaintance intends that the group

happens with the agent who starts the negotiation, whenever an agent holds an alternating intention that is the conclusion of its own practical inference, it will also intend to make all the other members know about its intention (see Proposition 4). In most cases, this is achieved by the agent's generating a counter-proposal either by making a modification of the initial proposal or by making a new proposal [51]. In either case, a new plan will be suggested that will be another candidate for being moved up to an agreed-upon plan status within the group. Expressing this formally:

Proposition 7. If a member of a group jointly committed to attaining a state of the world comes to believe that another member intends that the group performs a particular plan[11], and if it intends that the group performs a different plan, then it will intend to bring about a state where every member holds a belief about its own intention that the group performs the latter plan[12]:

$\_$ $[\forall\, gr, \forall\, a_i, a_j, a_l \in gr, a_i, a_l \forall\, e_h, e_k\, (k \neq h), \forall\, t$
$[J\text{-}COMM(gr, \phi)(t) \wedge Int(a_l, \Diamond\langle plan(gr, e_h, \phi)\rangle)(t) \wedge Bel\ (a_i, Int\ (a_l, \Diamond\langle plan(gr, e_h, \phi)\rangle))(t) \wedge Int(a_i, \Diamond\langle plan(gr, e_k, \phi)\rangle)(t)] \supset \exists\, t' \geq t$ s.t.
$Int(a_i, Bel\ (a_l, Int\ (a_i, \Diamond\langle plan(gr, e_k, \phi)\rangle)))(t') \wedge$
$Int(a_i, Bel\ (a_j, Int\ (a_i, \Diamond\langle plan(gr, e_k, \phi)\rangle)))\ (t')$

Informally, Proposition 7 states that, in a jointly committed group, if an agent is aware of the fact that it disagrees with another agent as to which plan the group should perform, then it will intend to make all the other members know about its own view. As expected, Proposition 7 replicates Proposition 4, in that a conclusion of a practical inference is transformed into an intention to impact upon the members' mental states until they generate a new belief. The only difference is that here the agent's mental state is enriched by a belief about an acquaintance's proposal, and disagreement triggers the generation of a counter-proposal. The process then iterates until either *one* of the agents involved in the negotiation withdraws or *all* the agents agree on which plan is to be performed collectively. In the latter case, negotiation is successful and the whole process ends up with an agreement that invokes a means (i.e., the performance of a plan) for achieving a state collectively. An agreement about a plan inherently rests on all the agents' sharing a *joint commitment* to jointly performing that plan. In turn, this joint commitment can be seen as generated from a state in which each single agent *individually intends* that the group acts according to that plan. Expressing this formally:

---

performs a particular plan. That is, reality may differ from perceived reality. For a discussion of this point, see Section 5.2.

[11] As with Proposition 6, we are not assuming that the agent believes correctly (Section 5.2).

[12] Again, we are assuming that the agent believes that its inferential conclusion will still be true at the (future) time point when everybody in the group knows about this conclusion.

Proposition 8. If all the members of a group jointly committed to attaining a state of the world intend that the group performs a particular plan, then a state will follow in which they will be jointly committed to performing that plan[13]:

$\_ \forall gr, \forall a_i \in gr, \forall e_h, \forall t$

$J\text{-}COMM(gr, \phi)(t) \wedge Int(a_i, \Diamond\langle plan(gr, e_h, \phi)\rangle)(t) \supset \exists t' \geq t$ s.t.

$J\text{-}COMM(gr, \langle plan(gr, e_h, \phi)\rangle)(t')$

It is important to highlight that sharing identical individual intentions does not imply that the agents share the same practical judgement, namely the same view as to the most appropriate plan to be performed (see Section 5.1 for a discussion). In most real-world situations, the agents' judgements are different depending on what beliefs and preference functions they hold. As implied by Proposition 3, different practical judgements bring about different intention-based inferential conclusions. The purpose of negotiation is to find out how different inferential conclusions may converge on a mutually acceptable plan, and therefore on a shared identical intention that the group performs that plan. Finding such a common solution will lead the agents to compromise with one another over the selection of the plan. Once they all agree on the plan, they will share identical intentions that the group *will* act in a certain manner; however, they may still hold different individual practical judgements as to how the group *should* act. This means that the outcome of negotiation is not based on a full consensus within the group as to the most satisfactory plan. In this view, the agreed-upon plan cannot be seen as logically inferred from an inferential process jointly carried out by the group. Rather, it is the result of a trade-off between sticking to one's practical judgement at the risk of jeopardising the negotiation on the one hand, and making concessions and compromises in order to increase the overall joint gains among the agents on the other [58]. The essence of negotiation lies in the way in which agents manage to solve this trade-off, which in turn is reflected in the interplay between proposals, counter-proposals, concessions, critiques, objections, forms of justifications, rejections, and withdrawals from the process. Such an interplay is the real key to understanding the path from which a system of socially and cognitively differentiated agents may ultimately converge into a system of socially and cognitively integrated ones.

5. Properties of the Model

[13] An agreement about a joint plan requires not only identical intentions as to which plan the group should perform, but also awareness within the group that this is the case. Group awareness can be achieved by iterative message passing within the group until each agent knows that all the others know (and so on *ad infinitum*) that each agent holds the same intention. For simplicity, in Proposition 8 we focused only on the convergence of intentions as a necessary condition for agreement. Also, for simplicity and clarity we abstracted over specific mechanisms of convergence of intentions, assuming that an agreement occurs when all the agents' intentions are identical. However, different social choice mechanisms (e.g. democratic voting mechanisms; appealing to authority) might well have been adopted.

In this section, some of the most interesting properties of the model will be examined and discussed. According to the classical axiomatic-deductive methodology for theory building [55], these properties will be introduced as theorems formally derived from the assumptions of the model. As logical consequences of a set of premises, properties are statements that are necessarily true whenever the premises are true. In this view, a theory consists of all the statements that are logical consequences of the set of premises, and therefore includes the premises themselves, because they are trivial consequences of themselves. This, therefore, completes the development of our theory of negotiation.

In deriving and discussing the following properties, we will attempt to make some steps towards two related purposes. First, by deriving properties on a proof-theoretic basis, we will show to what extent the logical language in which our model has been expressed can be used as a powerful tool for theorising about multi-agent systems. Logical formalisation provides a number of strict criteria for theories, such as consistency and soundness of argumentation, which are difficult to impose on their discursive counterparts. When formalising a theory with logic, contradictions can be resolved and the underlying argumentation structure can be made explicit and transparent [53]. This, in turn, will shed light on the descriptive and prescriptive power of the theory, and on its parsimony and coherence. Second, while using computational and logical tools to address some of the major problems occurring in pluralistic social systems, we are making a contribution towards a cross-fertilisation between mainstream DAI and traditional organization theory and research. On the one hand, many of the concepts and problems on which the organisational thought and research have long been focused have been neglected by most scholars interested in the use of computational and logic-based formalisms for reasoning about multi-agent systems. On the other, the logical tools and methods that are of major concern in mainstream computer science and DAI have been mainly underestimated in the organisational literature. This presents us with a fundamental challenge, namely to incorporate divergent methodologies, techniques and vocabularies of two related scholarly fields into a comprehensive and encompassing conceptual framework that is both analytically rigorous and empirically satisfactory.

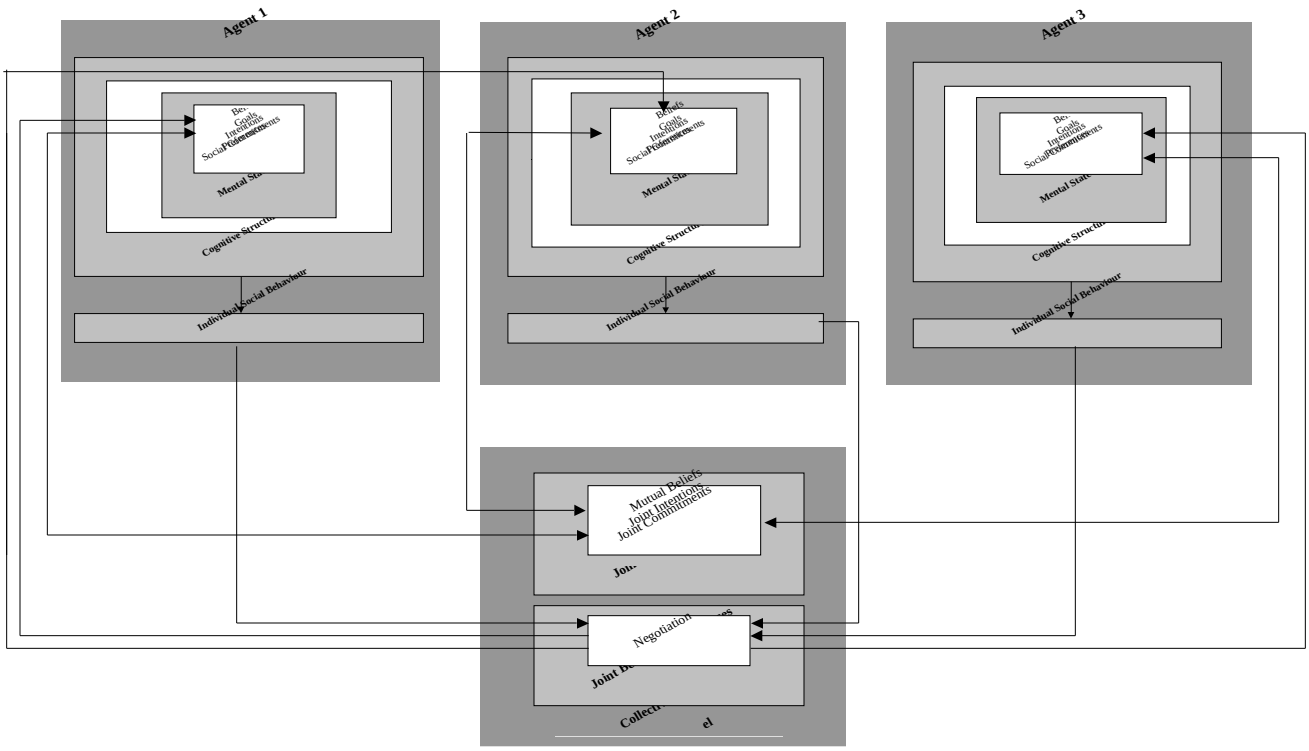
The remainder of this section is structured as follows. In Section 5.1 the concept of social and cognitive agenthood will be introduced and, building on that, properties will be formalized that relate the agreement to the individual agents' mental attitudes. Section 5.2 deals with the issue of the agents' bounded rationality and the related problem of informational inaccuracy that might affect the generation of an agreement. In Section 5.3, the problem of bounded rationality will be extended to a setting in which the agents' divergent interests and the related problem of opportunistic behaviour can be accounted for. Finally, Section 5.4 advocates a new contracting approach to modelling multi-agent systems, and introduces the notion of cognitive transaction as a promising means for reasoning about negotiation in a social setting.

5.1 Cognition at the individual and joint level

The objective of this section is to illustrate the role that cognition plays in negotiation, both at the individual and the joint level. To this end, we will first describe the concept of *socio-cognitive agenthood* that lies at the heart of our theory. Drawing on this conceptual framework, we will then use our logic to formalise and discuss some properties governing the interrelationship between agreement and individual agents' mental attitudes.

Figure 1 is intended to synthetically illustrate our notion of agenthood, its structure and the dynamic interplay of its main components. To this end, three agents engaged in a negotiation process are portrayed. On a more general level, agenthood is seen as existing at two interrelated levels: the individual and the collective one. At each level, agenthood is expressed in terms of cognition and behaviour. At the top of the figure, the individual level of cognition and behaviour is shown. At the bottom, the joint level is portrayed which rests on and at the same time impacts upon the individual one.

Figure 1. A Model of Negotiation: Relating Joint Behavioural Processes to Cognition at the Individual and Collective Levels.



Let us start our analysis with the individual level. Here, the negotiating agent is modelled as a cognitive agent capable of individual social behaviour [71]. For an agent to be termed a *cognitive agent* it

must be endowed with mental attitudes representing the world and motivating action (e.g. [49], [50], [59], [65], [73]). Further, for a cognitive agent to be a *socio-cognitive agent* it must not only have an intentional stance towards the world, but also represent other agents as cognitive agents similarly endowed with mental attitudes for representational and motivational purposes [24]. These properties of individual socio-cognitive agenthood can be conceptualised and operationalised by modelling the agent as having a *cognitive structure*. Crudely, according to mainstream cognitive science and expert systems research, the agent's cognitive structure can be defined as the set of basic mental attitudes, cognitive principles for modifying these attitudes, operators for adding mental attitudes, the cognitive frame, the language, control procedures, social interaction propensities, and principles and mechanisms for generating inter-agent social behaviour (e.g. [10], [23], [24], [32], [70]). The cognitive structure changes as the agent interacts with its physical and social environment and acquires new information. This, in turn, affects the agent's behaviour and its interaction with the environment [10]. A key role within the agent's cognitive structure is played by the *mental state*. This is the set of the basic mental attitudes that are processed by the agent to undertake both theoretical reasoning (i.e., reasoning undertaken to determine what is the case) and practical reasoning (i.e., reasoning undertaken to determine what is to be done) (e.g. [3], [49]). These mental attitudes include beliefs, goals, intentions, etc. (e.g. [19], [41], [59], [65]).

In the middle of Figure 1, the social behavioural dimension of the individual level is portrayed. Here, the agent is modelled as being capable of *individual social action* aimed at producing effects on its social environment. Specifically, a socio-cognitive agent's action is social when it is regulated by the agent's cognitive structure, and is oriented towards another agent who, in turn, is regarded as a cognitive agent whose behaviour is regulated by a cognitive structure [71]. According to this conception, the agent's individual social behaviour rests on the agent's cognitive structure, in that it is controlled by the principles of social behaviour which, in turn, are affected by the agent's mental state. Thus, the information the agent gathers and maintains about the world as well as the states of affairs it wants and/or intends to bring about are the ultimate causal forces on the top of the cognitive chain governing the agent's interaction with its social environment.

Moving onto the higher-order collective level. As shown at the bottom of Figure 1, the group comprising the three negotiating agents is represented. As with the individual agent, we have two fundamental components of the collective level: joint cognition and joint behaviour. On the one hand, the mental states of the members of the group are meshed together in such a way that a *joint mental state* will ensue (e.g., [49], [73]). This includes varying joint doxastic, motivational and deontic mental attitudes, such as mutual beliefs, joint intentions, joint commitments (e.g. [49], [73]). On the other hand, joint behaviour rests on and transcends the complex interplay between the agents' individual social behaviours. Joint mental state and joint behaviour co-evolve in that changes in one of them impact onto the other in a two-way

direction, through the mediation of the individual's cognition and behaviour. It is this complex web of "mediated" interrelationships that lies at the heart of our theory of negotiation.

Because of its key role in guiding negotiation, in what follows we will restrict our attention to the group's joint commitment. Our concern, therefore, will be to illustrate the role that individual cognition and behaviour play in governing the relationship between joint commitment and negotiation (joint behaviour). First, negotiation finds its cognitive roots in joint commitment, in a causal chain involving both individual social behaviour and cognition. It is a joint commitment to achieving some state that gets the committed agents' mental states to trigger and control the performance of individual social actions in order to reach an agreement about how to attain that state. Second, negotiation impacts back onto joint commitment by affecting the jointly committed agents' mental states. As the individual agent engages in the negotiation process, it may acquire new beliefs, fulfil its goals and intentions, or modify some of them. Experiential wisdom accumulates as a result of positive and negative reinforcement of prior mental attitudes [10]. Mental attitudes that have led to what are encoded as positive outcomes are reinforced, while those that have led to negative outcomes are modified or discharged [63]. Finally, as a result of its impact upon the agent's mental state, negotiation affects joint commitment as this is cognitively grounded on the individual agents' mental attitudes. For example, failure to reach an agreement may lead an agent to generate the belief that no agreement can reasonably be made. Or, negotiation may lead an agent to drop its motivation to carry on negotiation or to adopt an intention that is incompatible with its being involved in the process. In all these circumstances, the agent will exit negotiation and the joint commitment it maintained with the others will be dropped.

In the light of this conceptual framework, in what follows we will use our logic to formalise and discuss some properties concerning the relationship between agreement (i.e., joint commitment) and individual mental attitudes. First, on a more general basis, we note that at the heart of an agreement lie the individual agents' intentions. Should negotiation be successful, not only will the agents share an intention towards an end, but they will also share an intention towards the means to secure that end.

Property 1. Agreement rests on and transcends individual intentions. More formally, we have:

_ $\forall\, gr, \forall\, a_i, a_j \in gr, \forall\, t$

$[J\text{-}COMM(gr, \phi)(t) \wedge J\text{-}COMM(gr, <plan(gr, e_h, \phi)>)(t)] \supset$

$\forall\, a_i \in gr\ Int\ (a_i, \phi)\ (t) \wedge Int(a_i, \Diamond<plan(gr, e_h, \phi))\ (t)$

Proof. The proof is a straightforward application of our definition of joint commitment (see Section 3 and [49] for details).

Even though reaching an agreement implies endorsing identical intentions, nevertheless the agents' having identical intentions concerning a state and the joint performance of a plan to achieve that state does not necessarily imply that a negotiation has been carried out and an agreement has been reached (i.e., the implication above is unidirectional). The reason for this is quite straightforward. Having identical individual intentions is not enough for a joint commitment to take place (see definition of joint commitment). This has a two-fold implication. First, the conjunction of the agents' intentions towards a state ϕ does not imply a joint commitment towards ϕ, which, according to our model, is a pre-condition of negotiation. Second, in a similar way, the simple conjunction of the agents' intentions regarding the joint performance of some action e_h does not entail the group's joint commitment to performing e_h, which, in our formalisation, represents the ultimate outcome of negotiation.

The next property relates the agents' mental states with the content of a possible agreement among them. As shown with Property 1, agents who come to an agreement are expected to change their mental states until they share the same intention about a plan. In the simple case of a group of two agents, this may take place either when one of them adapts its mental state to the other's or when both change their mental states in a similar manner.

Property 2. Should two agents favour two different plans, agreement between them may converge upon either one of the two plans or a new different one. More formally, we have:

$$\_\ \forall gr, \forall a_i, a_j \in gr, \forall e_h, e_k, \forall t$$
$$[J\text{-}COMM(gr, \phi)(t) \wedge Int(a_i, \Diamond<plan(gr, e_h,\phi)>)(t) \wedge Int(a_j, \Diamond<plan(gr, e_k, \phi)>)(t)] \supset \exists\ t' \geq t \text{ s.t. } W([(Int(a_i, \Diamond<plan(gr, e_h, \phi)>) \wedge Int(a_j, \Diamond<plan(gr, e_h, \phi)>)) \vee (Int(a_i, \Diamond<plan(gr, e_k, \phi)>) \wedge Int(a_j, \Diamond<plan(gr, e_k, \phi)>)) \vee (\exists\ e_p\ (p \neq h,k) \text{ s.t. } Int(a_i, \Diamond<plan(gr, e_p, \phi)>) \wedge Int(a_j, \Diamond<plan(gr, e_p, \phi)>))], \neg [Comm(a_i, gr, \phi) \wedge Comm(a_j, gr, \phi)])(t')$$

Proof. For arbitrary gr, a_i, $a_j \in gr$, e_h, and t, and using Propositions 4 to 6 sequentially, we have: $J\text{-}COMM(gr, \phi)(t) \wedge Int(a_i, \Diamond<plan(gr, e_h,\phi)>)(t) \supset \exists\ t'' \geq t$ s.t. $Int(a_i, Bel(a_j, Int(a_i, \Diamond<plan(gr, e_h, \phi)>)))(t') \supset \exists\ t''' \geq t''$ s.t. $Bel(a_j, Int(a_i, \Diamond<plan(gr, e_h, \phi)>))(t''') \supset \exists\ t' \geq t'''$ s.t. $W([Int(a_j, \Diamond<plan(gr, e_h, \phi)>) \vee \exists\ e_p\ (p \neq j)$ s.t. $Int(a_j, \Diamond<plan(gr, e_p, \phi)>)], \neg Comm(a_j, gr, \phi))(t') \Leftrightarrow W([Int(a_j, \Diamond<plan(gr, e_h, \phi)>) \vee Int(a_j, \Diamond<plan(gr, e_{k(k \neq h)}, \phi)>) \vee \exists\ e_p\ (p \neq h, k)$ s.t. $Int(a_j, \Diamond<plan(gr, e_p, \phi)>)], \neg Comm(a_j, gr, \phi))(t')$. In a similar way, for arbitrary gr, a_i, $a_j \in gr$, e_k, and t, we obtain: $J\text{-}COMM(gr, \phi)(t) \wedge Int(a_j, \Diamond<plan(gr, e_k,\phi)>)(t) \supset \exists\ t'' \geq t$ s.t. $Int(a_j, Bel(a_i, Int(a_j, \Diamond<plan(gr, e_k, \phi)>)))(t') \supset \exists\ t''' \geq t''$ s.t. $Bel(a_i, Int(a_j, \Diamond<plan(gr, e_k, \phi)>))(t'') \supset \exists\ t' \geq t'''$ s.t. $W([Int(a_i, \Diamond<plan(gr, e_k, \phi)>) \vee \exists\ e_p\ (p \neq k)$ s.t. $Int(a_i, \Diamond<plan(gr, e_p, \phi)>)], \neg Comm(a_i, gr, \phi))(t')$

$\Leftrightarrow W([Int(a_i, \Diamond<plan(gr, e_k, \phi)>) \vee Int(a_j, \Diamond<plan(gr, e_{h(h \neq k)}, \phi)>) \vee \exists e_p \ (p \neq k, h)$ s.t. $Int(a_i, \Diamond<plan(gr, e_p, \phi)>)], \neg Comm(a_i, gr, \phi)) (t')$.

Informally, the above property means that two agents who differ in the plan they support, may give up their intentions and agree to support a new different plan. Because they are not fanatical, they will eventually give up negotiation unless one of them adopts the other's intention or both decide to endorse and share a new one. This property can be extended to the whole group of agents so that the agreement can be viewed as not limited to the agents' views and preferences. All the agents can change their mental states, give up their individual intentions and adopt a new one for the sake of an agreement.

Property 2 has an interesting implication that concerns one of the key problems found in real-world negotiations: *compromising* and *intention reconsideration*. According to Property 2, the agents may drop their individual intentions for the sake of the group (e.g. [13], [66], [69]). However, they do not drop their views and personal preferences. Even though they come to share an identical intention about the joint performance of a plan, the agents may still have differing preferences and personal views as to what is the most appropriate plan the group should perform. At the heart of this dichotomy between views/preferences and intentions lies the essence of compromising. By adopting a new intention, the agents compromise with each other over their own ideas. In turn, compromising brings about a process of intention reconsideration. Let us suppose that a group *gr* has successfully carried out a negotiation: $J\text{-}COMM(gr, \phi)(t) \wedge J\text{-}COMM(gr, <plan(gr, e_h, \phi)>)(t)$. Furthermore, suppose that one of the agents, say a_i, still maintains a practical judgement favouring a plan that is different from what has been agreed upon within the group: $Bel(a_i, (\Diamond\phi \Leftrightarrow \Diamond<plan(gr, e_k, \phi)>))(t) \vee \wedge_{p \neq k} Pref(a_i, plan(gr, e_k, \phi), plan(gr, e_p, \phi))(t)$. Now, according to Property 1, the agreement on e_h implies $Int(a_i, \Diamond<plan(gr, e_h, \phi))(t)$. However, according to Proposition 3, a practical judgement implies the generation of the corresponding intention, $Int(a_i, \Diamond<plan(gr, e_k, \phi))(t')$, where $t' \geq t$. If agent a_i is to be regarded as a rational agent [73], and if actions e_h and e_k cannot be performed simultaneously or are mutually exclusive, then a_i cannot hold the two intentions at the same time. Nor, if an agreement has been reached regarding the performance of action e_h, can the agent intend that the group performs either e_h or e_k.[14] In the light of this, an intention reconsideration must occur which leads the agent to drop the intention based on its own practical judgement in order to endorse the intention favoured by the whole group. Endorsing a socially motivated intention to the detriment of an internally motivated one represents the cognitive implication of the compromises that the agents are often required to make with one another over their own ideas in order to get to a final agreement.

[14] This is because the formula $Int(a_i, (\Diamond<plan(gr, e_h, \phi) \vee \Diamond<plan(gr, e_k, \phi))(t)$ is satisfied also when $Int(a_i, \Diamond<plan(gr, e_h, \phi))(t)$ is not satisfied and $Int(a_i, \Diamond<plan(gr, e_k, \phi))(t)$ is satisfied, whereas an agreement on e_h implies that $Int(a_i, \Diamond<plan(gr, e_h, \phi))(t)$ needs always to be satisfied.

However, compromising and intention reconsideration are not a general rule for negotiation. The agents may share the same views and preferences without being aware of this. They therefore need to negotiate in order to find out they all agree on what the group should do. In this case, no conflict exists between the agents' ideas and therefore no compromising is required. This gives us the opportunity to conclude this section with a couple of observations about the role of conflict in negotiation. It has been argued that negotiation is a response to conflict [27]. However, as noted above, in some circumstances negotiation is carried out in the absence of conflict. Instead, in order for conflict to occur, at least two agents need to hold two different views, and hence two different intentions as to how to fulfil their joint commitment. Furthermore, even though necessary, nevertheless two different intentions are not sufficient to trigger conflict. The two agents must be aware of each other's different intentions before they can engage in social interactions aimed at reconciling them. In fact, as long as the difference is not perceived, the agents will not be motivated to overcome it. "Who intends what" and "who believes who intends what" are therefore the two key ingredients of conflict and the social activity undertaken to manage it.

5.2 Private information, bounded rationality, and informational incompleteness and asymmetries.

A key problem in most real-world negotiations is the uncertainty and ambiguity of the information needed to reach an agreement and to determine whether the terms of it are mutually acceptable. There are two main reasons for this. First, in most circumstances different agents have differing relevant *private information* before an agreement is reached. As a result of this, the information that is needed to reach an agreement tends to be localised and dispersed throughout the multi-agent system. Second, in most real-world scenarios agents are *boundedly rational* (e.g., [62], [63], [64]). They have limited cognitive ability, imperfect communication skills and their natural languages are imprecise. As a result, agents cannot solve arbitrarily complex problems exactly, costlessly and instantaneously, they cannot process all the information they have simultaneously and accurately, they cannot communicate with one another freely and perfectly, and the understanding of messages is often flawed. In the light of this, determining who to interact with, what information a message conveys, what message should be forwarded, to whom, and using what method becomes an overwhelmingly large and complex problem. Because no one has the cognitive ability to make all these calculations needed to retrieve information, and because information is localised and dispersed, not all the relevant information needed to determine the best use of resources and the appropriate adaptations can be fully accounted for by the agents before an agreement can be reached. It is at the heart of the link between the condition of informational dispersion and the agents' bounded rationality that lies the problem of *informational inaccuracy and asymmetry*, which, in turn represents a major obstacle that interferes with the possibility of reaching a mutually beneficial agreement (e.g. [47], [63], [64]).

Our model of negotiation is rich and flexible enough to enable the agents' informational incompleteness to be represented and accounted for. The following property is precisely intended to give a formalisation of the fact that, in most real-word negotiations, the agents' beliefs about each other's mental attitudes are not deterministically accurate. They are not inevitably true in the same way as they are not inevitably false. In this respect, and on more epistemological grounds, Property 3 conveys the view that the agents' doxastic representations of the world are inherently ambiguous and uncertain as a result of the agents' limited cognitive ability to overcome the problem of informational localisation and dispersion.

Property 3. Agents are boundedly rational in generating and updating their beliefs about each other's intentions. More formally, we have:

$\_\ \forall\ gr, \forall\ a_i, a_j \in gr, \forall\ e_h, \forall\ t, t'\ (t'>t)$
$[J\text{-}COMM(gr, \phi)(t) \wedge Bel\ (a_j, Int\ (a_i, \Diamond<plan(gr, e_h, \phi)>))(t, t')\ ] \supset \exists\ t'' > t$ s.t.
$W\ ([Int\ (a_i, \Diamond<plan(gr, e_h, \phi)>) \vee \exists\ e_k\ (k \neq h)$ s.t. $Int\ (a_i, \Diamond<plan(gr, e_k, \phi)>)], \neg Comm(a_i, gr, \phi))\ (t'')$

Proof. We begin by showing that the consequent in the implication above is implied by the first conjunct alone in the antecedent. First, let us note that: $W\ ([Int\ (a_i, \Diamond<plan(gr, e_h, \phi)>) \vee \exists\ e_k\ (k \neq h)$ s.t. $Int\ (a_i, \Diamond<plan(gr, e_k, \phi)>)], \neg Comm(a_i, gr, \phi))\ (t'') \Leftrightarrow W([Int(a_i, \Diamond<plan(gr, e_h, \phi)>) \vee Int(a_i, \Diamond<plan(gr, e_{q\ (q\neq h)}, \phi)>) \vee \exists\ e_k\ (k \neq h, q)$ s.t. $Int(a_i, \Diamond<plan(gr, e_k, \phi)>)], \neg Comm(a_i, gr, \phi))\ (t'') \Leftrightarrow W([Int(a_i, \Diamond<plan(gr, e_h, \phi)>) \vee Int(a_i, \Diamond<plan(gr, e_{p\ (p\neq h)}, \phi)>) \vee \exists\ e_k\ (k \neq h, p)$ s.t. $Int(a_i, \Diamond<plan(gr, e_k, \phi)>)], \neg Comm(a_i, gr, \phi))\ (t'')$. Now, for arbitrary $gr, a_i, a_j \in gr, e_h$, and $t, t'\ (t'>t)$, and using Propositions 1 and 2 sequentially, we have: $J\text{-}COMM(gr, \phi)(t) \wedge Bel(a_j, Int\ (a_i, \Diamond<plan(gr, e_h, \phi)>))(t, t') \supset \exists\ t'''>t, a_l, e_q$ s.t. $[Bel(a_l, (\Diamond\phi \Leftrightarrow \Diamond<plan(gr, e_q, \phi)>))(t''') \vee \wedge_{p \neq q} Pref\ (a_l, plan(gr, e_q, \phi), plan(gr, e_p, \phi))\ (t''') \vee \exists\ e_p$ s.t. $\wedge_{p \neq q} Pref\ (a_l, plan(gr, e_p, \phi), plan(gr, e_q, \phi))\ (t''')]$. We now need to distinguish between two cases. First, let us suppose the following is satisfied: $Bel(a_l, (\Diamond\phi \Leftrightarrow \Diamond<plan(gr, e_q, \phi)>))(t''') \vee \wedge_{p \neq q} Pref\ (a_l, plan(gr, e_q, \phi), plan(gr, e_p, \phi))\ (t''')$. Using Propositions 3 to 6 sequentially, we have: $[Bel(a_l, (\Diamond\phi \Leftrightarrow \Diamond<plan(gr, e_q, \phi)>))(t''') \vee \wedge_{p \neq k} Pref\ (a_l, plan(gr, e_q, \phi), plan(gr, e_p, \phi))\ (t''')] \supset \exists\ t^{IV} \geq t'''$ s.t. $Int(a_l, \Diamond<plan(gr, e_q, \phi)>))(t^{IV}) \supset \exists\ t^{V} \geq t^{IV}$ s.t. $Int(a_l, Bel(a_i, Int(a_l, \Diamond<plan(gr, e_q, \phi)>)))(t^{V}) \supset \exists\ t^{VI} \geq t^{V}$ s.t. $Bel(a_i, Int(a_l, \Diamond<plan(gr, e_q, \phi)>))(t^{VI}) \supset \exists\ t'' > t^{VI}$ s.t. $W([Int(a_i, \Diamond<plan(gr, e_q, \phi)>) \vee \exists\ e_k\ (k \neq q)$ s.t. $Int(a_i, \Diamond<plan(gr, e_k, \phi)>)], \neg Comm(a_i, gr, \phi))\ (t'') \Leftrightarrow W([Int(a_i, \Diamond<plan(gr, e_q, \phi)>) \vee Int(a_i, \Diamond<plan(gr, e_{h\ (h\neq q)}, \phi)>) \vee \exists\ e_k\ (k \neq q, h)$ s.t. $Int(a_i, \Diamond<plan(gr, e_k, \phi)>)], \neg Comm(a_i, gr, \phi))\ (t'')$. Second, suppose the following is satisfied: $\exists\ e_p$ s.t. $\wedge_{p \neq q} Pref\ (a_l, plan(gr, e_p, \phi), plan(gr, e_q, \phi))\ (t''')$. Again, by applying Propositions 3 to 6, we obtain $\exists\ t'', t^{IV}, t^{V}, t^{VI}$ s.t. $(t''> t^{VI} \geq t^{V} \geq t^{IV} \geq t''' >t) \wedge W([Int(a_i, \Diamond<plan(gr, e_p, \phi)>) \vee \exists\ e_k\ (k \neq p)$ s.t. $Int(a_i, \Diamond<plan(gr, e_k, \phi)>)], \neg Comm(a_i, gr, \phi))\ (t'') \Leftrightarrow W([Int(a_i, \Diamond<plan(gr, e_p, \phi)>) \vee Int(a_i, \Diamond<plan(gr, e_{h\ (h\neq p)}, \phi)>) \vee \exists\ e_k\ (k \neq p, h)$ s.t. $Int(a_i, \Diamond<plan(gr,$

e_k, ϕ)>)], $\neg$*Comm*(a_i, *gr*, ϕ)) (*t''*). Thus, in either case, we obtained that the first conjunct alone of the implication entails the consequent. As to the second conjunct, it is straightforward to observe that our model contains no axiom or proposition in which what is implied by an agent's belief contradicts what is implied by the joint commitment of the group to which the agent belongs. This completes the proof.

Informally, Property 3 means that, should an agent have a belief about another's intention, a state will follow in which the latter will drop its commitment to the group unless it endorses an intention concerning the joint performance of a plan. However, this intention need not be the same as the one represented in the former's mental state. Therefore, there is no guarantee that an agent's belief about another's intention is accurate. More specifically, Property 3 allows for two complementary conceptualisations of the agent's bounded rationality. First, when $t'' = t'$, Property 3 means that the agent's beliefs about what is currently the case may be inaccurate. In fact, the agent may mistakenly believe that another holds an intention. The main reason for this is that the agent typically uses imperfect cognitive representations to form mental models of its environment [67]. These representations thus simplify the complexity of both spatial [56] and temporal or causal relationships [72]. Second, when $t'' < t'$, Property 3 means that, not only may the agent be inaccurate in generating its beliefs, but it also may be unable to update its beliefs once formed. In fact, when $t'' < t'$, the agent may keep maintaining a belief that another holds an intention, even though the latter does not hold that intention. Here, the agent's limited ability to update its beliefs does not allow it to discover every change in its environment that might falsify its cognitive representations formed at an earlier stage. Even beliefs that were accurately formed may subsequently turn out to be false during the course of negotiation as a result of changes in the environment.

In the light of this, the problem of informational incompleteness and asymmetries can be extended to a dynamic setting in which the inability to obtain all the relevant information also arises from the agents' inability to foresee and unambiguously describe every contingency that might possibly be relevant to an agreement. With bounded rationality, contingencies will arise that have not been accounted for because they were never imagined at an earlier stage (e.g. [62], [63], [64]). This might occur as a result not only of the agents' limited foresight, but also of their inability to solve arbitrarily complex problems exactly, costlessly and instantaneously. In fact, even when contingencies can be foreseen, they may appear so unlikely that it is not considered worth investing resources in describing them in detail. This is most likely the case when the opportunity costs of the agents' time spent foreseeing things rather than doing productive activities are high, or when the contingencies seem not to be detrimental to the negotiation process should they occur. Yet, even in all these cases, all the calculations needed to foresee circumstances are subject to error, thus generating inaccuracy in the agents' belief set.

5.3 Conflicting interests and opportunistic behaviour

The property that will be discussed in this section is concerned with another central problem of negotiation among boundedly rational agents: *opportunistic behaviour* and the related issue of *motivation*. Not only have real agents limited cognitive ability. They also have their own private interests, which are rarely perfectly aligned with the interests of the other agents with whom they need to interact (e.g. [20], [21]). Divergence of interests, together with bounded rationality and information specificity, introduces the possibility of opportunistic behaviour. Because agents are boundedly rational, they suffer from informational distortions. Not only can this cognitive weakness per se prevent the parties from reaching a mutually beneficial agreement. It can also be exploited by the agents to opportunistically misrepresent or even refuse to reveal relevant private information in order to obtain a unilateral advantage and seize a greater share of the fruits of negotiation for themselves [48]. Correspondingly, the motivation problem is to ensure that the various agents involved in negotiation willingly do their parts in the whole undertaking, both communicating information accurately to allow the right agreement to be reached and acting as they are expected to act within the group.

The issue of agents' divergent objectives and interests is a delicate matter. Among European studies, Pettigrew's [54] analysis of the politics of organisational decision-making was one of the earliest and most influential works criticising Simon's model for its exclusive focus on the problem of informational uncertainty posed by the agents' bounded rationality. In his detailed study of a decision concerning the introduction of a central computer in a firm, Pettigrew showed how the lack of relevant and accurate information interacted with and was intentionally exploited to accomplish the conflicting interests of the various agents involved. When the emphasis is on the agents’ conflicting interests, negotiation can thus be regarded as precisely an integrative mechanism capable of governing pluralistic multi-agent systems (e.g. [21], [22], [61]). Not only does negotiation regulate the exchange and pooling of agents' resources, interests and actions, but it also defines and reshapes the governing structure, the "rules of the game". In this sense, as a mode of regulating pluralistic multi-agent systems, negotiation differs from other integrative mechanisms such as democracy (a system of governance of the many integrated by voting mechanisms; see [42], [45]) and anarchy (a governance system of extreme ambiguity mainly characterised by a lack of knowledge and clarity about whom the agents are and what their preferences are; see [18]). As contrasted with these alternative forms of integration, negotiation is a more goal-oriented mode of governing a pluralistic system in which equilibrium and efficiency are highly dependent on the interplay between the agents' potentially conflicting interests.

In this view, a key motivation problem in a negotiation-based system becomes one of arranging affairs so that, as far as possible, an agent's individual social behaviour takes proper account not only of how that agent is affected by an agreement, but of how others are affected as well. In fact, should the agents not be sufficiently motivated to act in a way that is beneficial to the whole group, they might behave opportunistically and hide relevant private information, or even alter it in an effort to have their own interests

and objectives satisfied at the expense of the others'. The "market for lemons" case is an eloquent example [1]. As has been argued, in the used car market, sellers have better information about their cars than potential buyers do. Fear that prices are not fair probably sours many a worthwhile deal. This source of inefficiency is often called *adverse selection*, conveying the idea that one party's offers (in the example above, the selection of cars in the market) are determined in a way that is detrimental, adverse to the interests of the other party.

It is precisely the problem of adverse selection that we will pinpoint with the following property. On a proof-theoretic basis, it can be shown how our model is flexible enough to allow the agents' opportunistic behaviour, and in particular the problem of adverse selection, to be accounted for. Not only are the agents in the real world boundedly rational nor perfectly capable of cognitively representing their environment, but they also know they have such imperfections. Particularly, they recognise their acquaintances cannot possibly have all the information that might matter for them, nor can they always acquire it. This opens the possibility of self-interested misbehaviour aimed at misrepresenting relevant private information during negotiation [48]. An agent may intend that another perceives that it has a specific intention concerning a potential agreement, perhaps because this may induce the latter to act in a manner that is beneficial to the former. Or, it may well be the case that an agent intends to hide its own strategy to another agent, thus deliberately forwarding wrong messages to the latter in an attempt to make it generate inaccurate beliefs.

Property 4. Agents may opportunistically mislead each other into thinking that they maintain intentions they actually do not. More formally, we have:

$\_\ \forall\ gr, \forall\ a_i, a_j \in gr, \forall\ e_h, \forall\ t, t'\ (t'>t)$
$[J\text{-}COMM(gr, \phi)(t) \wedge Int(a_i, Bel\ (a_j, Int\ (a_i, \Diamond<plan(gr, e_h, \phi)>)))\ (t, t')] \supset \exists\ t'' > t$ s.t.
$W\ ([Int\ (a_i, \Diamond<plan(gr, e_h, \phi)>) \vee \exists\ e_k, k \neq h,$ s.t. $Int\ (a_i, \Diamond<plan(gr, e_k, \phi)>)], \neg Comm(a_i, gr, \phi))\ (t'')$

Proof. We need to show first, that the first conjunct alone of the implication entails the consequent; second, that the second conjunct has no implication that contradicts what is implied by the first conjunct. The proof is a straightforward adaptation of the earlier proof of Property 3.

Informally, Property 4 means that an agent's intention that another believes that it has an intention concerning a plan does not entail that it really favours that plan. In fact, should that agent maintain its commitment to the group, it will eventually endorse an intention that may well favour a different plan. As it stands, "who intends who believes who intends what" does not affect "who intends what". This result is similar to the one shown in Property 3, where an agent's belief may turn out to be either true or false, having no cogent guarantee underpinning its accuracy. Along these lines, it can be argued that our model has been inspired by an epistemological diffidence towards determinism, in that the agents are allowed to be

mistakenly led to shape their minds and form their mental attitudes. In this respect, the model conveys a view combining a cognitive orientation, a concern for the cognitive freedom of the agents, and an awareness of the impossibility of substantive deterministic predictions of cognitive processes and behaviours.

Before we move onto other properties of our model, there is another important aspect of the problem of divergent interests and opportunistic behaviour that needs to be highlighted here. So far, we have glossed over the topic of self-interested misbehaviour that might affect the achievement of a mutually beneficial agreement between two or more agents. This is a form of *pre-contractual* opportunism that arises because of the private information that boundedly rational agents have *before* they reach an agreement. Besides this, there is another form of self-interested misbehaviour that occurs after an agreement is made. Our model is centred around the cognitive and social properties of the transformation of a prior joint commitment into another derivative joint commitment. Although our main concern is about the generation of an agreement which is reflected in the latter derivative commitment, the model is implicitly based on an underpinning form of meta-agreement: this is an agreement about reaching an agreement. In fact, being jointly committed to negotiating as to how to attain some state of affairs means to agree to attain that state in a collaborative manner, and therefore to agree to eventually make an agreement about the appropriate means to secure the end. This allows us to briefly discuss that form of *post-contractual* opportunism known as *moral hazard* that arises *after* an agreement has been made [52].

At the root of post-contractual opportunism lie the dynamic implications of the agents' bounded rationality. Not only have agents inaccurate information about the world; they are also not perfectly far-sighted [63]. In particular, agents' ability to make agreements are limited by the existence of unforeseen circumstances, the costs of deciding in advance what would be appropriate to do in every foreseeable contingency, and the difficulty to obtain information about contingencies with enough precision to make the search for such information worthwhile. Even in the extreme case, where there is no private information before an agreement is made, there may be inadequate information afterward to tell whether the terms of the agreement have been honoured, or acquiring that information may be costly. As a result of this, agreements are incomplete and imperfectly specified, so that the agents involved can exploit loopholes to gain an advantage over one another [52]. In addition, as shown with the problem of adverse selection, actions that have efficiency consequences are not freely observable and so the agent taking them may choose to pursue its private interests at others' expense [48].

In the light of this, Property 4 can be read from a different perspective that enables the problem of moral hazard to be accounted for. Because boundedly rational agents cannot foresee all the relevant circumstances, the agreements they make to negotiate are inevitably imperfect and incomplete. This is true also of a meta-agreement concerning how to reach an agreement. That is, the prior joint commitment to collaboratively engaging in negotiation cannot conceivably specify all the relevant circumstances that might arise during negotiation. This opens the incentive for agents with divergent interests to opportunistically

misbehave in such a way that their own private interests can be fulfilled with no mutual advantage for the others. Again, opportunistic misbehaviour is captured in our model by an agent's intending that another mistakenly perceives something, after a joint commitment has been formed that binds the two agents to negotiate. As the way in which negotiation is to be carried out cannot be perfectly and completely specified in advance, the agents, once a joint commitment is formed, are induced to misbehave and try to fulfil the joint commitment in a self-interested manner.

5.4 The cognitive transaction

According to one of the major approaches to the economic analysis of multi-agent systems, the ability to enter and undertake negotiations and contracts is critical for reaching individual and collective goals [2]. In this view, a multi-agent system can be regarded as a *nexus* of agreements, contracts, treaties, and understandings among the individual members. Using this approach, the most fundamental unit of analysis is the transaction, namely the transfer of good or services from one individual to another [76]. Along these lines, it can be conjectured that our model builds on and extends the contracting approach to multi-agent systems to a setting in which agreements provide the cognitive foundations for pluralistic social systems. In fact, the typical function of agreements is to allow cognitively differentiated agents to become cognitively integrated. In this account, the *cognitive transaction* represents the key unit of analysis. This form of transaction can be defined as the transfer of mental attitudes from an agent to another [50]. More specifically, in order for a cognitive transaction to take place, at least two cognitive agents are needed. Additionally, one of the two agents must take an intentional stance towards the other [24]. Given this, a cognitive transaction occurs when an agent is motivated by its cognitive representation of another's mental attitude to adopt that attitude. For example, the adoption by a member of a political group of the leader's goal is an instantiation of a cognitive transaction. In this view, a cognitive transaction does not need the two agents' mutual awareness of each other. Instead, it simply requires that an agent modifies and re-shapes its mental state so as to include another's mental attitude.

In the light of the above observations, a cognitive transaction is inherently cognitive for a two-fold reason. First, it is fundamentally grounded on an agent's taking an intentional stance towards another [24]. Second, it involves the exchange of epistemic, doxastic, conative, and deontic attitudes that are intrinsically rooted in the agents' mental states and cognitive processes. In addition, a cognitive transaction is also inherently social because the exchange of mental attitudes is carried out in a social setting of differentiated agents, and is typically intertwined with inter-agent social behaviour[15] [71].

[15] However, not all cognitive transactions occur within social relationships. For instance, an agent may adopt another's belief or goal as a consequence of the latter's charisma or authority. An example is the case of a political leader who may cause a diffusion of his own views, ideals, goals within its sphere of influence without being aware of the identity of each individual who contributes to the diffusion of such attitudes by adopting them.

In order for a cognitive transaction to take place, at least two cognitive agents are needed. Additionally, one of the two agents must take an intentional stance towards the other [24]. Given this, a cognitive transaction occurs when an agent is motivated by its cognitive representation of another's mental attitude to adopt that attitude. For instance, In this view, a cognitive transaction does not need the two agent's mutual awareness of each other. Instead, it simply requires that an agent modifies and re-shapes its mental state so as to include another's mental attitude.

The variety and complexity of the ways of organising negotiations found in the real-world reflect some of the following basic attributes of cognitive transactions:

1. The *specificity* and *opaqueness* of mental attitudes. Mental attitudes are specific cognitive assets that are not freely accessible from outside the agents' mental states. In addition, in real-world scenarios agents are boundedly rational and use imperfect cognitive representations of their environment. This is true also when agents' cognitive representations are about others' cognitive representations. As a result of the combined effect of these two related problems (cognitive specificity/opaqueness and bounded rationality), agents' beliefs about others' mental attitudes are not necessarily accurate. As pointed out with the issue of private information (Property 3), the truth of "who believes who intends what" does not affect the truth of "who intends what". Along these lines, it can be conjectured that an agent's maintaining a mental attitude as a result of its believing that someone else has that mental attitude is not necessarily the outcome of a cognitive transaction. This is because should that agent's belief be false, no mental attitude would be actually transferred from an agent's mental state to another's. More specifically, in order to give a formal account of the specificity/opaqueness of mental attitudes with respect to a cognitive transaction, we have the following property:

Property 5. Given a jointly committed group, a member's belief about another's intention is a necessary but not sufficient condition for a cognitive transaction to occur between the two agents.

The formalisation and proof of this property are identical to Property 3, and we therefore omit them for simplicity. In essence, the meaning of the property is that an agent's believing that someone else has an intention does not entail the truth of that intention, and therefore is not enough to trigger a cognitive transaction. In fact, should the agent's belief be false, its adopting the intention cannot be truly described as a transfer of a mental attitude from the other agent's mental state to its own.

2. The *uncertainty* about the performance. Uncertainty about the conditions that will prevail when a cognitive transaction is being executed together with the complexity of the task makes it impossible or at least ambiguous to determine in advance what the outcome will be. Even in the unrealistic case in which

mental attitudes are not specific cognitive assets and agents are perfectly rational, an agent's forming a mental representation of another's mental attitude cannot guarantee a successful execution of a cognitive transaction. There two main reasons for this. First, autonomous agents have cognitive freedom, and in most cases can decide whether or not to adopt another's mental attitude [50]. Second, a cognitive transaction is inherently connected to many forms of inter-agent social behaviour. In a social setting, agents are often the subjects of complex networks of social relationships and, even outside relationships, are variously subjected to the influence of others. They strategically use this networks of relationships and influence to shape their own mental states by finding creative interpretations of others' attitudes [10]. This makes the evaluation of a cognitive transaction inherently uncertain and highly dependent on exogenous social factors and chance events. In terms of our model, and within the context of negotiation, this uncertainty can be expressed with the following property.

Property 6. Given a jointly committed group, a member's intention as to how to fulfil the joint commitment, and another's belief about the former's intention are a necessary but not sufficient condition for a cognitive transaction to occur between the two agents. More formally, we have:

$\_\ \forall\ gr, \forall\ a_i, a_j \in gr, \forall\ e_h, \forall\ t$

$[J\text{-}COMM(gr, \phi)(t) \wedge Bel\ (a_j, Int\ (a_i, \Diamond<plan(gr, e_h, \phi)>))(t) \wedge Int\ (a_i, \Diamond<plan(gr, e_h, \phi)>))(t)\ ] \supset \exists\ t' > t$ s.t.

$W\ ([Int\ (a_j, \Diamond<plan(gr, e_h, \phi)>) \vee \exists\ e_k\ (k \neq h)$ s.t. $Int\ (a_j, \Diamond<plan(gr, e_k, \phi)>)], \neg Comm(a_j, gr, \phi))\ (t')$

Proof. We need to show that the consequent is implied by the first conjunct alone, without the other two conjuncts having a contradictory implication. The proof is a straightforward adaptation of earlier results. In our model, we have no proposition in which the last two conjuncts above are assumed to have an implication that contradicts the consequent of Property 4. In addition, we have already shown that a joint commitment alone entails that consequent.

Informally, the meaning of the implication above is that an agent's belief about another's intentions, even in the case in which it is accurate (i.e. the other agent truly maintains the intention), may not results in the former agent's adopting the latter's intention. An agent's intending and agent's believing about the other's intending are the necessary conditions in order for a cognitive transfer of the intention to occur from the latter's mental state to the former's. However, they are not sufficient to guarantee that transfer. Because their combined effect upon the agent's mental state is highly dependent on the agent's ultimate decision and other social and chance factors, it cannot be deterministically represented. This, in turn, does not enable a cognitive transaction to be deterministically predicted.

3. The *difficulty of measuring* the performance. The uncertainty and unpredictability of a cognitive transaction is strictly related to the *difficulty of measuring* its performance, that is the problem of evaluating whether and to what extent a cognitive transaction has occurred. The main reason for this is that it is highly uncertain to determine what the cognitive source of an agent's mental attitude is. An attitude may be either socially or internally motivated [50], and the way in which it is generated does not deterministically affect its content, structure and role it plays in the agent's mental state. Furthermore, even in the case in which the *desired* performance of a cognitive transaction is predictable (e.g., an agent intends to adopt another's mental attitude), it may be difficult or costly to measure its actual performance. For example, the agent may mistakenly believe that a cognitive transaction has occurred, whereas the adoption of the mental attitude may have been driven by internal motives or even different socio-cognitive sources (i.e. different attitudes of different agents). In terms of our model, this property can be expressed in the following way.

Property 7. Given a jointly committed group, a member's intention as to how to fulfil the joint commitment cannot necessarily be viewed as the result of a cognitive transaction that occurred between two agents. More formally, we have:

$\_\ \forall\ gr, \forall\ a_i \in gr, \forall\ e_h, \forall\ t, t'(t' < t)$
$[J\text{-}COMM(gr, \phi)(t', t) \wedge Int\ (a_i, \Diamond<plan(gr, e_h, \phi)>))(t)] \supset \exists\ t'' > t', \exists\ a_j \in gr$ s.t.
$[Bel(a_i, Int\ (a_j, \Diamond<plan(gr, e_h, \phi)>))(t'') \wedge Int\ (a_j, \Diamond<plan(gr, e_h, \phi)>))(t'')] \vee \exists\ e_k\ (k \neq h)$ s.t. $[Bel(a_i, Int\ (a_j, \Diamond<plan(gr, e_k, \phi)>))(t'') \wedge Int\ (a_j, \Diamond<plan(gr, e_k, \phi)>))(t'')]$

Proof. First, note that the following holds: $[Bel(a_i, Int\ (a_j, \Diamond<plan(gr, e_h, \phi)>))(t'') \wedge Int\ (a_j, \Diamond<plan(gr, e_h, \phi)>))(t'')] \vee \exists\ e_k\ (k \neq h)$ s.t. $[Bel(a_i, Int\ (a_j, \Diamond<plan(gr, e_k, \phi)>))(t'') \wedge Int\ (a_j, \Diamond<plan(gr, e_k, \phi)>))(t'')] \Leftrightarrow [Int(a_j, \Diamond<plan(gr, e_h, \phi)>)(t'') \wedge Bel\ (a_i, Int\ (a_j, \Diamond<plan(gr, e_h, \phi)>))\ (t'')] \vee [Int(a_j, \Diamond<plan(gr, e_{p\ (p \neq h)}, \phi)>)(t'') \wedge Bel\ (a_i, Int\ (a_j, \Diamond<plan(gr, e_{p(p \neq h)}, \phi)>))\ (t'')] \vee [Int(a_j, \Diamond<plan(gr, e_{q\ (q \neq h, p)}, \phi)>)(t'') \wedge Bel\ (a_i, Int\ (a_j, \Diamond<plan(gr, e_{q(q \neq h, p)}, \phi)>))\ (t'')] \vee \exists\ e_k\ (k \neq h, p, q)$ s.t. $[Int\ (a_j, \Diamond<plan(gr, e_k, \phi)>))(t'') \wedge Bel(a_i, Int\ (a_j, \Diamond<plan(gr, e_k, \phi)>))(t'')]$. Second, note that when $J\text{-}COMM(gr, \phi)$ is satisfied in the interval (t', t), it is also satisfied in t'. Now, applying Propositions 1 to 5 sequentially, we have: $\forall\ t', \forall\ a_i \in gr\ J\text{-}COMM(gr, \phi)(t') \supset \exists\ t'' > t', \exists a_j, \in gr, \exists e_p$ s.t. $[Int(a_j, \Diamond<plan(gr, e_p, \phi)>)(t'') \wedge Bel\ (a_i, Int\ (a_j, \Diamond<plan(gr, e_p, \phi)>))\ (t'')\} \vee \exists\ e_q\ (q \neq p)$ s.t. $[Int\ (a_j, \Diamond<plan(gr, e_q, \phi)>))(t'') \wedge Bel(a_i, Int\ (a_j, \Diamond<plan(gr, e_q, \phi)>))(t'')] \Leftrightarrow [Int(a_j, \Diamond<plan(gr, e_p, \phi)>)(t'') \wedge Bel\ (a_i, Int\ (a_j, \Diamond<plan(gr, e_p, \phi)>))\ (t'')] \vee [Int(a_j, \Diamond<plan(gr, e_{h\ (h \neq p)}, \phi)>)(t'') \wedge Bel\ (a_i, Int\ (a_j, \Diamond<plan(gr, e_{h(h \neq p)}, \phi)>))\ (t'')] \vee [Int(a_j, \Diamond<plan(gr, e_{k\ (k \neq p, h)}, \phi)>)(t'') \wedge Bel\ (a_i, Int\ (a_j, \Diamond<plan(gr, e_{k(k \neq p, h)}, \phi)>))\ (t'')] \vee \exists\ e_q\ (q \neq p, h, k)$ s.t. $[Int\ (a_j, \Diamond<plan(gr, e_q, \phi)>))(t'') \wedge Bel(a_i, Int\ (a_j, \Diamond<plan(gr, e_q, \phi)>))(t'')]$.

Finally, there is no proposition in the model according to which the implication of $Int(a_i, \Diamond<plan(gr, e_h, \phi)>))(t)$ contradicts what is implied by $J\text{-}COMM(gr, \phi)(t', t)$, even when $t = t''$.

Informally, when $t'' \leq t$, the implication in Property 7 conveys the idea that, during negotiation, an agent's intention may not be the consequence of its believing that someone else has that intention. Negotiating agents' intentions as to how to fulfil the joint commitment are not necessarily socially motivated in the same way as they are not necessarily internally motivated. The fact that there is a fundamental ambiguity about the source of mental attitudes makes it difficult to deterministically specify whether they have been adopted as a result of a cognitive transaction. From this perspective, it can be argued that Property 7 complements property 6, in that it completes the picture of the agent's cognitive freedom. In this view, the agent is the ultimate depository of the decision whether or not to adopt a mental attitude. Not only is it uncertain whether any set of conditions can bring about the agent's adoption of a mental attitude (Property 6), but it is also uncertain to determine whether an attitude has been adopted for certain reasons (Property 7). That is, the agent's mental state cannot be deterministically predicted in the same way as it cannot be deterministically explained.

The properties discussed so far in this section highlighted the fact that the cognitive transaction is an amorphous and ambiguous notion that is reluctant to be placed in a nomological network of antecedents and consequences. However, cognitive transactions are crucial to negotiation. To a certain extent, it can be argued that the way in which a negotiation is organised and evolves reflects the structure and evolution of the nexus of cognitive transactions that occur between the agents involved. Because negotiation is typically concerned with the problem of transforming socially and cognitively differentiated agents into integrated ones, the cognitive transaction represents the most straightforward means through which socio-cognitive integration can be achieved. In fact, adopting another's mental attitude is the first step towards the achievement of a shared cognitive basis on which, in turn, an agreement will ultimately be generated (e.g. [43], [73], [75]). Along these lines, the complexity that can be found in most real-world negotiations can be regarded as the mirror of the complex and ambiguous nature of the underpinning cognitive transactions and their interplay. Therefore, understanding the properties of these transactions not only can bring explanatory power to theories of negotiation, but can also offer invaluable tools for the problem of managing and designing efficient ways in which agreements are reached in the real world.

6. Conclusions and discussion

We started our work by posing a challenge, namely the possibility of integrating divergent analytical tools, techniques, principles, and research questions from different theoretical perspectives into a unified

paradigm. This paper took some steps towards meeting this challenge by developing a theoretical framework for modelling the cognitive foundations for pluralism in multi-agent systems. By addressing the problem of how to conceptualise and formalise the negotiation process among intelligent agents, we set out to justify the claim that drawing upon the interconnections among various scientific disciplines interested in MAS has the potential to significantly improve our understanding of pluralistic forms of organising social action. At the heart of our claim lies a new view of MAS that is endorsed by recent advances in social networks, cognitive sciences, and DAI - the idea of MAS as computational, complex and adaptive systems in which knowledge and action occur at multiple levels (e.g. [39], [73]).

This view has motivated our attempt towards a two-pronged integration of theoretical perspectives. Firstly, we brought some of the major research questions in social sciences to bear on the methods and analytical tools advocated by mainstream computer science and DAI. In this respect, we attempted to formalise such problems as the agent's bounded rationality, pre- and post-contractual opportunistic behaviour, using a computational BDI logic, and the axiomatic-deductive methodology for developing argumentations. Secondly, we worked towards a cross-fertilisation among research questions by bridging the gap between different units of analysis that are endorsed by different theoretical perspectives to study the same object. In doing this, we studied for instance the problem of informational asymmetries and incompleteness, typically addressed by organisational scientists, by focusing on the agent's cognitive architecture, which is a conceptual category imported from DAI.

Our methodological and theoretical integration of perspectives allowed us to gain new insights into the core problems of pluralistic intelligent behaviour. In fact, the theory of negotiation we developed is rooted in a new conception of agenthood that, in turn, is sympathetic with the above mentioned view of multi-agent systems as computational systems in which knowledge and cognition are embedded in multiple levels. According to this new conception of agenthood, the individual agent is regarded as a cognitive and social entity, endowed with an intentional stance and capable of reasoning about other agents in terms of their intentional stance [24]. This view enables sociality to be represented in the agent's mind and the emergence of higher-level group mind-like constructs from the interaction among agents to be accounted for. In this vein, negotiation has been modelled as governed not only by the individual agent's cognition, but also by the group's joint mental state that rests on and transcends the individuals' mental attitudes. Finally, from an epistemological viewpoint, our notion of agenthood conveys the idea of the agent as a cognitively free actor, capable and allowed to decide whether or not to adopt mental attitudes, to join and construct organisational arrangements, and to perform individual social behaviour. An eloquent example of the agent's cognitive freedom is our notion of cognitive transaction that has been built on the idea that the agent's mental state cannot be deterministically predicted in the same way as it cannot be deterministically explained. Finally, even though we have been reluctant to adopt a deterministic conception of cognition, we have nevertheless endorsed a view that is also diffident towards the agent's *cognitive anarchism* and

complete indeterminism. In fact, our axiomatic-deductive methodology for theory building enabled us to model and derive high-level principles for governing the structure and evolution of the agent's cognitive architecture at both the individual and joint level. Even though they are not intended to prescriptively indicate how human agents should reason, these principles can nevertheless be regarded as a specification for the design of autonomous artificial agents. In this vein, our theory can be used to derive a set of data structures, algorithms and principles that may be particularly suitable for developing cooperative intelligent systems of artificial negotiating agents. In this respect, besides its main theoretical and methodological contributions, our work is intended to provide assistance to practitioners who are primarily concerned with building distributed computer systems to be used to accomplish multi-agent tasks and solve real-word problems in a wide range of social, industrial and commercial applications.

References

[1] G. Akerlof, The market for lemons: Qualitative uncertainty and the market mechanism, Quarterly Journal of Economics 84 (1970) 488-500.

[2] A. Alchian and H. Demsetz, Production, information costs, and economic organization, American Economic Review 62 (1972) 777-795.

[3] R. Audi, A theory of practical reasoning, American Philosophical Quarterly 19(1) (1982) 25-39.

[4] S. B. Bacharach and E. J. Lawler, Power and Politics in Organisations: The Social Psychology of Conflict, Coalitions, and Bargaining, Jossey-Bass, San Francisco, 1980

[5] J. Bell. Changing attitudes, in: *Intelligent Agents, Post-Proceedings of the ECAI-94 Workshop on Agent Theories, Architectures, and Languages*, eds. M. Wooldridge and N. R. Jennings, Springer, Berlin, 1995, pp. 40-55.

[6] Bell and Z. Huang, Dynamic goal hierarchies, in: *Intelligent Agent Systems: Theoretical and Practical Issues*, eds. L. Cavedon, A. Rao & W. Wobcke, Springer-Verlag, 1997, pp. 88-103.

[7] K. Binmore, Modeling rational players I & II, Economics and Philosophy 3 (1987) 9-55 & 4 (1987) 179-214.

[8] A. H. Bond and L. Gasser, editors. *Readings in Distributed Artificial Intelligence*, Kaufmann, San Mateo, California, 1988.

[9] R.M. Burton and B. Obel, The validity of computational models in organisation science: From model realism to purpose of the model, Computational and Mathematical Organisation Theory 1(1) (1995) 57-71.

[10] K. Carley, The value of cognitive foundations for dynamic social theory, Journal of Mathematical Sociology, 14(2-3) (1989) 171-208.

[11] K. Carley, A theory of group stability, American Sociological Review 56(3) (1991) 331-354.

[12] K. Carley, On the evolution of social and organizational networks, in: Networks In and Around Organizations, Vol. 16 Special Issue of Research in the Sociology of Organizations, eds. S. B. Andrews and D. Knoke, JAI Press, Inc. Stamford, CT, 1999, pp. 3-30.

[13] K. Carley and A. Newell, The nature of the social agent, Journal of Mathematical Sociology 19(4) (1994) 221-262.

[14] K. Carley and M. J. Prietula (eds.), *Computational Organisation Theory*, Hillsdale, NJ, Lawrence Erlbaum Associates, 1994.

[15] K. Carley and W. A. Wallace, Editorial, Computational and Mathematical Organisation Theory 1(1) (1995) 5-7.

[16] B. Chellas. *Modal Logic: An Introduction*, Cambridge University Press, 1980.

[17] D. S. Clarke. *Practical Inferences*, Routledge and Kegan Paul, London, 1985.

[18] M. D. Cohen, J. J. March and J. P. Olsen, A garbage can model of organizational choice, Administrative Science Quarterly 17 (1972) 1-25.

[19] P. R. Cohen and H. J. Levesque, Intention is choice with commitment, Artificial Intelligence, 42(3), (1990) 213-261.

[20] M. Crozier, *Le Phénomène Burocratique*, Paris, Éditions du Seuil, 1963.

[21] M. Crozier and E. Friedberg, *L'Acteur et le Système*, Paris, Éditions du Seuil, 1977.

[22] M. Crozier and J.-C. Thoenig, La regulation des systemes organises complexes, Revue Francaise de Sociologie, 16(1) (1975) 3-32.

[23] Davis, R., and D. Lenat, *Knowledgebase Systems in Artificial Intelligence*, New York, NY, McGraw Hill, 1980.

[24] D. C. Dennet. *The Intentional Stance*, The MIT Press, Cambridge, MA, 1987.

[25] P. Faratin, C. Sierra and N. R. Jennings, Negotiation decision functions for autonomous agents, Robotics and Autonomous Systems 24(3-4) (1998) 159-182.

[26] K. Fisher, J. J. Mueller, M. Pischel, and D. Schier, A model for cooperative transportation scheduling, in: *Proceedings of the International Conference on Multi-Agent Systems (ICMAS-95)*, AAAI/MIT Press, 1995, pp. 109-116.

[27] L. Greenhalgh and D. I. Chapman, Joint decision-making. The inseparability of relationships and negotiation, in: *Negotiation as a Social Process*, eds. R. M. Kramer and D. M. Messick, Thousand Oaks, CA, SAGE Publications, 1995, pp. 166-185.

[28] B. Grosz and S. Kraus. Collaborative plans for complex group actions, Artificial Intelligence 86 (1996) 269-357.

[29] R. Grunder. On the actions of social groups, Inquiry 19 (1976) 443-454.

[30] J. Y. Halpern and Y. Moses, A guide to completeness and complexity for modal logics of knowledge and belief, Artificial Intelligence 54 (1992) 319-379.

[31] D. Harel, Dynamic Logic, in: *Handbook of Philosophical Logic Volume II – Extensions of Classical Logic* , eds. D. Gabbay and F. Guenther, D. Reidel Publishing Company, Dordrecht, The Netherlands (Synthese library Volume 164), 1984, 497-604.

[32] F. Hayes-Roth, D. A. Waterman and D. R. Lenat (eds.), *Building Expert Systems*, Reading, MA, Addison-Wesley, 1983..

[33] J. Hintikka. *Knowledge and Belief*, Cornell University Press, 1962.

[34] J. Hintikka. Semantics for propositional attitudes, in *Reference and Modality*, ed. L Linsky, Oxford University Press, 1972.

[35] M. Hughes and M.J. Cresswell, *Introduction to Modal Logic*, Metheun and Co. Ltd., 1968.

[36] N. R. Jennings, P. Faratin, M. J. Johnson, T. J. Norman, P. O'Brien and M. E. Wiegand, Agent-based business process management, International Journal of Cooperative Information Systems 5(2&3) (1996) 105-130.

[37] N. R. Jennings, K. Sycara and M. Wooldridge, A roadmap of agent research and development, Autonomous Agents and Multi-Agent Systems 1 (1998) 275-306.

[38] D. Kinny, M. Ljungberg, A. S. Rao, E. Sonenberg, G. Tidhar and E. Werner, Planned team activity, in: *Artificial Social Systems – Selected Papers from the Fourth European Workshop on Modelling Autonomous Agents in a Multi-Agent World,* eds. C. Castelfranchi and E. Werner, *MAAMAW-92*, Vol. 830, Springer-Verlag, Heidelberg, Germany, 1992, pp. 226-256.

[39] D. Krackhardt and K. Carley, A PECANS model of structure in organizations, in: *Proceedings of the International Symposium on Command and Control Research and Technology*, Monterey, CA, 1997.

[40] S. Kraus and D. Lehmann, Designing and building an automated negotiating agent, Computational Intelligence 11(1), (1995) 132-171.

[41] S. Kraus, K, Sycara and A. Evenchil, Reaching agreements through argumentation: A logical model and implementation, Artificial Intelligence 104 (1998) 1-69.

[42] C. J. Lammers, Interorganizational democracy, in: *Interdisciplinary Perspectives on Organization Studies*, eds. S. Lindenberg and H. Schreuder, Pergamon Press, pp. 323-337.

[43] Levesque, J., P. R. Cohen, and J. H. T. Nunes, On acting together, in: *Proceedings of the Eighth National Conference on Artificial Intelligence, AAAI-90*, 1990, pp. 94-99.

[44] B. Levitt and J. G. March. Organisational learning, Annual Review of Sociology, Palo Alto, CA, Annual Reviews 14 (1988) 319-340.

[45] S. M. Lipset, M. A. Trow and J. S. Coleman, *Union Democracy*, New York, Free Press, 1956.

[46] R. D. Luce and H. Raiffa, *Games and Decisions*, New York, Dover, 1989; also appeared as: Utility theory, in: *Rationality in Action. Contemporary Approaches*, ed. P.K. Moser, New York, NY, Cambridge University Press, 1990.

[47] R. D. Middlemist and M. A. Hitt, *Organisational Behaviour, Managerial Strategies for Performance*, West Publishing Company, 1988.

[48] P. Milgrom and J. Roberts, Bargaining costs, influence costs, and the organization of economic activity, in: *Perspectives on Positive Political Economy*, eds. J. Alt and K. Schepsle, Cambridge, Cambridge University Press, 1990.

[49] P. Panzarasa, N. R. Jennings, and T. Norman, Formalising collaborative decision-making and practical reasoning in multi-agent systems, Journal of Logic and Computation (2001) forthcoming.

[50] P. Panzarasa, Jennings, N.R., and Norman, Social mental shaping: Modelling the impact of sociality on the mental states of autonomous agents, Computational Intelligence (2001) forthcoming.

[51] S. Parsons, C. Sierra and N. R. Jennings, Agents that reason and negotiate by arguing, Journal of Logic and Computation 8(3) (1998) 261-292.

[52] M. Pauly, The economics of moral hazard, American Economic Review 58 (1968) 31-58.

[53] G. Péli, J. Bruggeman, M. Masuch and B. Ó Nualláin, A logical approach to formalizing organizational ecology, American Sociological Review 59 (1994) 571-593.

[54] J. F. Pettigrew, *The Politics of Organization Decision Making*, London, Tavistock, 1973.

[55] K. R. Popper, *The Logic of Scientific Discovery*, London, Hutchinson, 1959.

[56] J. F. Porac, H. Thomas and C. Baden-Fuller, Competitive groups are cognitive communities: The case of Scottish knitwear manufacturers, Journal of Management Studies 26 (1989) 397-414.

[57] D.H. Pruitt, *Negotiation Behavior*, New York, NY, Academic Press, 1981.

[58] H. Raiffa, *The Art and Science of Negotiation*, Harvard University Press, Cambridge, MA, 1982.

[59] S. Rao and M. P. Georgeff, Modeling agents within a BDI architecture, in: *Proceeding of the Second International Conference on Principles of Knowledge Representation and Reasoning (KR '91)*, eds. R. Fikes, and E. Sandewall, Cambridge, MA, Morgan Kaufmann, 1991, pp. 473-484.

[60] A. S. Rao and M. P. Georgeff, Decision procedures for BDI logics, Journal of Logic and Computation 8(3) (1998) 293-342.

[61] F. W. Scharpf (ed), *Games in Hierarchies and Networks: Analytical and Empirical Approaches to the Study of Governance Institution*, Boulder, CO, Westview Press, 1993.

[62] H. A. Simon, A behavioural model of rational choice, Quarterly Journal of Economics 69 (1955) 99-118.

[63] H.A. Simon, *Administrative Behavior*, 3rd Edition. New York, NY, Free Press, 1976.

[64] H. A. Simon, Bounded rationality and organizational learning, Organization Science 2 (1991) 125-134.

[65] M. P. Singh. *Multiagent Systems: A Theoretical Framework for Intentions, Know-how, and Communications*, Springer-Verlag, Lecture Notes in Computer Science, Vol. 799, 1995.

[66] A. Strauss, Summary, implications, and debate, in: *Negotiations: Varieties, Contexts, Processes, and Social Order*, Jossey-Bass, San Francisco, 1978, pp. 234-262.

[67] P. Taghard, *Mind: Introduction to Cognitive Science*, Cambridge, MA, The MIT Press, 1996.

[68] M. Tambe, Towards flexible teamwork, Journal of Artificial Intelligence Research 7 (1997) 83-124.

[69] R. Tuomela, *A Theory of Social Action*, Reidel Pub., Boston, 1984.

[70] D.A. Waterman, *A Guide to Expert Systems*, Reading, MA, Addison-Wesley Publishing Co., 1986.

[71] M. Weber, *Economy and Society*, ed. G. Roth and C. Wittich, 2 Vols, Berkeley, CA, University of California Press, 1978.

[72] K. Weick, *The Social Psychology of Organizing*, second edition, Reading, MA, Addison-Wesley, 1979.

[73] M. Wooldridge, *Reasoning About Rational Agents*, Cambridge, MA, The MIT Press, 2000.

[74] M. Wooldridge and N. R. Jennings, Intelligent agents: Theory and practice, The Knowledge Engineering Review 10(2) (1995) 115-152.

[75] M. Wooldridge and N. R. Jennings, Cooperative problem solving, Journal of Logic and Computation 9(4) (1999) 563-594.

[76] O.E. Williamson, The economics of organization: The transaction cost approach, American Journal of Sociology 87 (1981) 548-575.

[77] G. H. von Wright. *The Logic of Preference*, Edinburgh, 1963.