The Cooperative Problem Solving Process

Michael Wooldridge and Nicholas R. Jennings
Journal of Logic & Computation, 9(4), pages 563-592, 1999

Abstract

We present a model of cooperative problem solving that describes the process from its beginning, with some agent recognising the potential for cooperation with respect to one of its goals, through to team action. Our approach is to characterise the mental states of the agents that leads them to solicit, and take part in, cooperative action. The model is formalised by expressing it as a theory in a quantified multi-modal logic.

1 Introduction

Agents — both human and artificial — can engage in many and varied types of social interaction, ranging from altruistic cooperation through to open conflict. However, perhaps the paradigm example of social interaction is *cooperative problem solving* (CPS), in which a group of autonomous agents choose to work together to achieve a common goal. For example, we might find a group of people working together to move a heavy object, play a symphony, build a house, or write a joint paper. In short, the aim of this paper is to develop a formal model of such cooperative problem solving.

Researchers working with the tools of game and economic theory have developed a number of models that attempt to explain various aspects of the cooperative problem solving process. Relevant examples include the circumstances under which cooperation can occur in a society of self-interested autonomous agents [1] and how negotiation protocols can be designed to ensure that (for example) truth-telling is the optimal strategy [23]. However, these models typically make assumptions that render them of limited value in many practical situations (we discuss the limitations of game-theoretic models in more detail in section 2.2). One of our aims in this article is, therefore, to present a formal model of CPS that is inherently more suitable as a *computational* model — we elaborate on this issue in section 7. Moreover, we wish the model to be *comprehensive*, in that it should cover the entire CPS process — from recognition of the need for cooperation through to completed team action. In more detail, the model consists of four stages:

- *recognition* — in which an agent identifies the potential for cooperation;
- *team formation* — in which the agent solicits assistance;
- *plan formation* — in which the newly formed collective attempts to construct an agreed joint plan; and finally,
- *execution* — in which members of the collective play out the roles they have negotiated.

The remainder of this article is structured as follows. Section 2 introduces the idea of CPS by way of some simple motivating examples. In order to formally express

this model, a new quantified multi-modal logic had to be devised, for representing the beliefs, goals, and actions of agents and groups of agents. This logic is informally introduced in section 3 (a complete formal definition of its syntax and semantics is given in Appendix A). The logic is used to formalise the notions of conventions, commitments, and intentions in section 4. These definitions are subsequently used in our model of CPS, which is presented in section 5. Section 6 discusses the properties of the model, and some conclusions and open issues are presented in section 7.

2 Modelling CPS: Issues and Scope

Many aspects of CPS have been investigated by researchers from distributed artificial intelligence, economics, philosophy, organisation science, and the social sciences. These models can be divided into two broad categories:

- implementation-oriented models for realising cooperative systems, managing cooperative activities, and achieving coordination in cooperative systems at run-time [27, 7]; and
- formal theories of cooperation and related issues; examples include economic and game-theoretic models of cooperation and negotiation [23], formal models of communication based on speech act theory (e.g., [6]), and models which typically use a multi-modal logic to describe the mental state of agents engaged in social activities [18, 22, 11].

Implementation-oriented models are useful in that they help to identify the various steps of the CPS process. For example, consider the Contract Net protocol [27]. This protocol contains the following steps: (i) *task announcement*: an agent (the manager) finds it has a problem that it does not have the resources to solve locally, and broadcasts an announcement to this effect; (ii) *bidding*: those agents that receive the announcement, and have the appropriate skills to help, send a *bid* to the manager, representing an offer to help; (iii) *awarding*: the manager awards the task to the most appropriate bidder, thus establishing a *manager-contractor* relationship between the two agents; and finally, (iv) *expediting*: the contractor carries out the task it has been awarded (which may involve generating sub-tasks, and further, hierarchical manager-contractor relationships). On completion of the task, the contractor informs the manager of the final result.

On examination, the Contract Net protocol reveals the following stages of CPS: (i) there is a point at which a manager recognises the potential for social action; (ii) there is an announcement stage, during which the prospective manager attempts to solicit assistance with respect to the task; (iii) there is a negotiation stage, during which potential managers and contractors engage in a dialogue, with the aim of agreeing which agent will do what; and (iv) there is a subsequent execution stage, during which participants play out the roles they have negotiated. Examination of other models of CPS (e.g., partial global planning [7]) indicates the same basic stages. Given this commonality, these are the four stages that our CPS model must cover.

Now that we have a broad understanding of the key steps that appear to be common to most forms of CPS, we can begin to identify the key properties that our model must satisfy. Such desiderata are presented in section 2.1. We then go on to

discuss the *purpose* of our model, and discuss why formalisation in symbolic logic is appropriate in section 2.2. In section 2.3, we discuss the different perspectives that such a model may take, and justify our choice of an *internal* perspective.

2.1 Desiderata for a Theory of Cooperative Problem Solving

We can identify the following desiderata for an adequate theory of the cooperative problem solving process:

- *Agents are autonomous.*

  Perhaps the most important requirement for a theory of CPS is that it cannot require *benevolence* (i.e., an *a priori* disposition to be helpful) on the part of agents. Agents are autonomous problem solvers [33]: hence they will take part in cooperative activities only if they *choose* to do so. A theory that simply required agents to cooperate whenever they were asked to would not be adequate, because it would fail to capture a significant proportion of real-world examples of cooperative activity.

- *Cooperation can fail.*

  A corollary of the fact that agents are autonomous is that cooperation may fail. If agents are not *required* to cooperate, then sometimes they won't. Even when initial cooperation is established, it can subsequently fail for many different reasons. For example, a group of agents that agree to cooperate in principle may discover that the assumptions upon which their choices were made do not in fact hold. Alternatively, events beyond the control of the team may make successful completion of their cooperation impossible. An adequate theory of cooperation must recognise that such failure is possible, identify the key points at which it may occur, and characterise the behaviour of a rational agent in such circumstances.

- *Communication is essential.*

  Although something resembling cooperation *is* possible *without* communication [28], we argue that communication is so fundamental to the everyday process of cooperation that an adequate theory should describe when and where communication should take place. That is, it should *predict* communication.

- *Communicative acts are characterised by their effect.*

  An adequate theory of cooperation should not *prescribe* the means through which communication actually takes place. For example, one possibility in a formal theory would be to define a number of message types, and assume that communication takes place by exchanging such messages. This would be unsatisfactory, however, as each of us makes use of many different methods for communication in our everyday lives, ranging from formal written statements and instructions to entirely informal and personal devices. All such approaches are equally valid from the CPS perspective.

- *Agents initiate social processes.*

Cooperation does not arise from a vacuum. It occurs because a group of agents believe they will in some way benefit from it. For example, an agent might believe that the cooperative solution to a problem is in some way better than a non-cooperative one: it may be more accurate or more up-to-date, for example. An adequate theory of cooperation should account for both the circumstances under which agents will begin to initiate cooperation and when they will initiate the social processes required to instantiate and complete cooperative actions.

- *Agents will be mutually supportive.*

  Cooperating agents will support one another during the execution of their joint action [3]. By this, we mean that agents will execute their part of the team's action, and will typically do what they can to ensure that the remainder of the team does likewise. An adequate theory of cooperation must describe the types of mutual support, when it should occur, and what form such support should take.

- *Agents are reactive.*

  Any realistic environment is highly dynamic. Agents must recognise this, and respond accordingly to any changes that affect their plans [33]. An adequate theory of cooperation must therefore recognise this reactive aspect of rational behaviour, and characterise the behaviour of the agents in such circumstances.

2.2 The Role of our CPS Model

Formalism in AI and multi-agent systems research plays many roles, which are too often confused. When presenting a logical theory of cooperative problem solving, it is therefore important to be precise about the role we expect the theory to play. In general, logical theories in multi-agent systems play one (or more) of the following roles: they can be exercises in *formal philosophy*, attempting to capture the properties of some human social activity in a precise way; they can be *specifications* for future computer systems, which attempt to prescribe the way in which a rational, intelligent system should behave; or they can be *knowledge representation* formalisms, intended to be directly represented and manipulated within some system.

Our theory is primarily intended as a *specification* for future cooperative systems. We have taken a number of extant models of cooperation and cooperative activity and from them abstracted the common components. We have then formalised this model in a multi-modal logic. The model we have derived cannot be implemented directly, since the modal logic we use to express the model does not lend itself to direct execution. (Direct execution of a logical formula corresponds to a constructive proof of satisfiability for that formula. Even for propositional multi-modal logics of the type we consider in this paper, the satisfiability problem is extremely complex [13].)

However, we argue that the model can be used to derive a set of data structures and algorithms that may be used to realise a cooperative system. We comment on this issue further in section 7.

Why not Game Theory?

Game and economic theory has proved to be one of the most successful formalisms for understanding cooperative behaviour [23]. It helps us to understand the parameters of cooperation, how it can arise, under what circumstances it is likely to succeed, and so on. However, while game theory is a useful *analytical* tool, it is not generally a good *engineering* tool, with which to build computational systems. This is primarily because of the *type* of representation employed by game theory. The building block of game theory is the notion of utility, whereby agents are allocated a real-valued *payoff* for every outcome in a particular encounter. A 'rational' agent is then one that acts to maximise its expected payoff in such an encounter.

Such simple abstractions lead to powerful models, that have been used to great effect in analysing the way an 'ideal' agent would behave in a multi-agent encounter [2]. However, game theoretic models are recognised to be idealisations of the way that agents would operate: they are not *computational* models, and ignore the practicalities of *computing* an appropriate action to perform [24]. Moreover, assuming the presence of a utility function, which assigns payoffs to possible outcomes, is simply not practicable for many real-world problems. In this sense, game-theoretic models are simply too coarse-grained for direct implementation in real systems. For these reasons, we choose to express out model of CPS as a logical theory.

2.3 Components and Perspectives

When devising a model of cooperative activity, we must choose a *perspective* from which to view the activity. There are two alternatives: external and internal [26]. With an *external* perspective, the actions performed by the agents are studied in order to determine when and how well the agents are cooperating; with an *internal* perspective, the agent's internal state is used as the basis for evaluation. For the reasons described below, this work adopts an *internal* perspective.

The first reason for using an internal approach is that it provides a high-level specification tool for the designer of a cooperating agent — it identifies the agent's key data structures, defines the relationships which exist between these structures, and places some constraints on the values which the structures can take (see [17] for an illustration of how an internal perspective model of cooperation was used to derive the high-level architecture of a social agent). The prescriptive nature of this approach contrasts with external models, which are mainly concerned with developing theories *about* agents, rather than on models which might be used *by* agents. (To reiterate, we are *not* suggesting that the logic be used directly as a knowledge representation formalism or programming language.)

The second reason for adopting an internal approach is that with the external perspective, it is sometimes difficult to distinguish between actions that are coordinated, but which one would not be inclined to call cooperative, and actions that are truly cooperative, in that the participating agents have a collective goal. To illustrate this point, consider the following scenario [25]. A group of people are sitting in a park. As a result of a sudden downpour all of them run to a tree in the middle of the park because it is the only available source of shelter. This is not cooperative action. Each person has the intention of stopping themselves from becoming wet, and even if they are aware of what others are doing and what their goals are, it does not affect their

intended action. This contrasts with the situation in which the people are dancers, and the choreography calls for them to converge on a common point (the tree). In this case, the individuals are performing exactly the same actions as before, but because they are performing these actions as a consequence of a shared goal, they can be regarded as performing a cooperative action. The external approach is not able to distinguish between the individuals trying to stay dry, and the cooperating dancers.

Having fixed upon an internal perspective, the next stage is to identify and characterise the structures which control an agent's cooperative problem solving activities. These structures can be divided into two categories: (i) those related to individual behaviour, and (ii) those that are responsible for guiding social behaviour. A number of researchers believe that joint action can be reduced solely to individual mental states; whereas others believe that individual behaviour is equivalent to social behaviour in which the groups have precisely one element. Our CPS model requires both individual and societal features to be present. Group constructs (such as teams, joint goals, joint commitments, and so on) are a natural tool for describing social activity; however, since it is the individuals who ultimately have the ability to act, there must be a clear mapping to the individual mental states of the participating agents. We therefore define social attitudes in terms of individual attitudes: following [5, 18], we take individual beliefs and goals to be primitive, and define other constructs, including those which characterise collective mental states, in terms of them.

3 A Formal Framework

This section gives an overview of the formal framework in which the model of CPS will be expressed; a complete formal definition is given in Appendix A. This framework is a quantified, sorted multi-modal logic, which both draws upon and extends the work described in [5, 22, 30]. The logic can be viewed as the well-known branching time logic CTL* [8], enriched by the addition of some further modal connectives for referring to the *beliefs* and *goals* of agents, together with a simple apparatus for representing the actions performed by agents, which makes use of some ideas from dynamic logic [14].

First, it is worth saying a few words about the *models* that underpin the logic. Intuitively, a model is a *time tree*, with paths through the tree representing possible histories of the environment. The tree will be finite in the past (i.e, there was a 'start' of time), and infinite in the future (i.e., there is no 'end' of time). Time is linear in the past, and branches into the future. Nodes in the temporal tree structure are referred to as *states*, for they correspond to states of the environment. Arcs in the branching time structure are labelled with *primitive actions*. The performance of such an action transforms one state into another. (We do not require that actions are deterministic.) Each primitive action is associated with a single agent, that performs the action.

The logic is quantified and many-sorted; for simplicity, we do not allow functional terms in the language other than constants. Terms come in four sorts. First, we have terms that denote *agents*, and we use $i, j, \ldots$ and so on as variables ranging over agents. In addition, we have terms that denote *sets* of agents, i.e., groups — we use $g, g', \ldots$ as variables ranging over groups of agents. Next, we have terms that denote *sequences of actions* — we use $\alpha, \alpha', \ldots$ as terms denoting sequences of actions. The role that such terms play will become clear later. Finally, we have terms that denote

other objects in the environment — bits of string, cars, tables, blocks, and so on. We will use $a, b, \ldots$ to stand for such individuals. The actual logical apparatus of quantification is standard for quantified many-sorted logics.

The logic makes a distinction between formulae that express properties of states, and formulae that express properties of paths, or histories through the temporal tree structure. The former are known as *state formulae*, the latter as *path formulae*. We begin our introduction by discussing the various *state formulae* operators (see Table 1 for an overview of the state and path operators in the logic). First, we have a nullary operator true: a logical constant for truth. This formula will be satisfied wherever it is evaluated. Next, we have operators $(\mathsf{Bel}\ i\ \varphi)$ and $(\mathsf{Goal}\ i\ \varphi)$, which mean that agent i has a belief and goal of φ respectively. An agent's beliefs intuitively correspond to the *information* that the agent has about its environment. For example, an agent might believe that the temperature of the room is 20 degrees celcius, or that Bill Clinton is a liar. Agents can have *nested* beliefs; thus an agent might believe that Bill Clinton did not believe of himself that he was a liar. For technical reasons, we require that an agent only believes state formulae. The formal semantics for belief are given in terms of 'possible worlds' [4]. The restrictions to be imposed on the language model theory ensure a belief logic of KD45, which thus implies that belief is consistent and closed under implication, and that an agent is aware of what it does and does not believe. The modal system KD45 is widely recognised as a logic of idealised belief [12].

Turning to goals, the idea is that an agent's goals represent those states of affairs that, ideally, it would like to bring about. For example, an agent might have a goal that the temperature in the room be 20 degrees celcius, or might have a goal that Bill Clinton be impeached. As with beliefs, an agent's goals must be state formulae, and the semantics of goals are also given in terms of possible worlds. Restrictions on the semantics of goals ensure that the logic of goals corresponds to a modal logic KD, i.e., the modal system D [4]. Thus goals are closed under implication, and are consistent.

In addition to these two modal connectives, we have first-order equality: a formula $(\tau = \tau')$ will be true if τ and τ' denote the same individual. The $\in$ operator allows us to relate agents to groups of agents. It has the expected set-theoretic interpretation, so $(i \in g)$ means that the agent denoted by i is a member of the group denoted by g. In order to allow us to represent the agents required to perform a sequence of actions, we have an operator Agts. This operator takes two arguments: the first is a term denoting a sequence of actions, the second is a term denoting a set of agents. Thus $(\mathsf{Agts}\ \alpha\ g)$ means that the group denoted by g are precisely the agents required to perform the actions in the action sequence denoted by α. We allow state formulae to be combined using the usual connectives of classical logic: '$\neg$' for 'not', '$\vee$' for 'or', '$\wedge$' (and), '$\Rightarrow$' (implies), and so on.

We now consider path formulae. As we noted above, the idea is that path formulae express properties of a single path through a branching time structure. The main operator for expressing the properties of paths is 'Happens'. This operator takes a single argument: an *action expression*, and expresses the fact that this action expression is the first thing that happens on the path. Action expressions closely resemble the programs of dynamic logic, so the path formula $(\mathsf{Happens}\ \alpha)$ will be satisfied on some path if the program α is the first thing to occur on the path.

Action expressions are formed using constructions that are well-known from dynamic logic: ';' (for sequential composition), '|' (for non-deterministic choice), '$*$' (for

(Bel i φ)	agent i believes φ
(Goal i φ)	agent i has a goal of φ
$(\tau = \tau')$	term τ denotes the same as τ'
$(i \in g)$	agent i is a members of group g
(Agts α g)	group g is required to do action sequence α
Aφ	on all paths, φ holds (inevitably φ)
(Happens α)	action expression α happens next

Table 1: A summary of primitive operators in the logic

iteration), and '?' (for test actions). Thus the path formula (Happens $\alpha; \alpha'$) means action α happens first on the path, and is immediately followed by α'. The formula (Happens $\alpha|\alpha'$) means either α or α' happen first on the path. The formula (Happens $\alpha*$) means that the action α occurs one or more times at the start of the path. Finally, the formula (Happens $\varphi?$) means that the formula φ is satisfied in the first state of the path. Here, φ must be a state formula. As with state formulae, compound path formulae can be made by combining path formulae using the standard logical connectives '$\neg$' for 'not', '$\vee$' for 'or', and so on.

State and path formulae are related to one another through *path quantifiers*, a concept borrowed from branching temporal logic [8]. The logic contains two such path quantifiers: 'A', which means 'on all paths', and 'E', which means 'on some path'. These path quantifiers are unary modal connectives that are applied to path formulae to make state formulae. Thus Aφ is a state formula, which will be satisfied in some state if the path formula φ is satisfied on all the paths through the temporal tree structure that originate from that state. The formula Eφ is a state formula, which will be satisfied in some state if φ is satisfied on at least one path through the temporal tree structure that originates from that state.

3.1 Some Derived Operators

A number of derived operators will now be introduced. First, the usual connectives of linear temporal logic: $\varphi \mathcal{U} \psi$ means φ is satisfied *until* ψ becomes satisfied; $\Diamond\varphi$ means φ is *eventually* satisfied; $\Box\varphi$ means φ is *always* satisfied. These connectives are used to build path formulae. The path quantifier E is the dual of A; thus Eφ means φ is a path formulae satisfied on *at least one* possible future.

$$\varphi \mathcal{U} \psi \stackrel{\text{def}}{=} (\text{Happens } (\neg\psi?; \varphi?)*; \psi?) \qquad \Box\varphi \stackrel{\text{def}}{=} \neg\Diamond\neg\varphi$$
$$\Diamond\varphi \stackrel{\text{def}}{=} \text{true}\,\mathcal{U}\,\varphi \qquad \text{E}\varphi \stackrel{\text{def}}{=} \neg\text{A}\neg\varphi$$

The next derived operators allow us to relate agents and groups of agents. The operators $\subseteq$ and $\subset$ relate groups together, and have the obvious set-theoretic interpretation; (Singleton g i) means g is a singleton group with i as the only member; (Singleton g)

simply means g is a singleton.

$$\begin{array}{rcl}(g \subseteq g') & \stackrel{\text{def}}{=} & \forall i \cdot (i \in g) \Rightarrow (i \in g') \\ (g \subset g') & \stackrel{\text{def}}{=} & (g \subseteq g') \wedge \neg(g = g') \\ (\mathsf{Singleton}\ g\ i) & \stackrel{\text{def}}{=} & \forall j \cdot (j \in g) \Rightarrow (j = i) \\ (\mathsf{Singleton}\ g) & \stackrel{\text{def}}{=} & \exists i \cdot (\mathsf{Singleton}\ g\ i)\end{array}$$

$(\mathsf{Agt}\ \alpha\ i)$ means i is the only agent of α.

$$(\mathsf{Agt}\ \alpha\ i) \stackrel{\text{def}}{=} \forall g \cdot (\mathsf{Agts}\ \alpha\ g) \Rightarrow (\mathsf{Singleton}\ g\ i)$$

To capture the notion of an action α *achieving* a goal φ, we introduce a derived operator Achieves :

$$(\mathsf{Achieves}\ \alpha\ \varphi) \stackrel{\text{def}}{=} \mathsf{E}(\mathsf{Happens}\ \alpha) \wedge \mathsf{A}((\mathsf{Happens}\ \alpha) \Rightarrow (\mathsf{Happens}\ \alpha; \varphi?))$$

Thus $(\mathsf{Achieves}\ \alpha\ \varphi)$ is very similar to the dynamic logic $\langle\alpha\rangle\mathsf{true} \wedge [\alpha]\varphi$. Thus not only does $(\mathsf{Achieves}\ \alpha\ \varphi)$ indicate that if α happens, then φ is true afterwards, but also that α does indeed occur on some path [14].

We will have a number of occasions to write $\mathsf{A}(\mathsf{Happens}\ \alpha)$, (action α occurs next in all alternative futures), and $\mathsf{A}\neg(\mathsf{Happens}\ \alpha)$ (action α does not occur next in any alternative future), and so we introduce abbreviations for these structures.

$$\begin{array}{rcl}(\mathsf{Does}\ \alpha) & \stackrel{\text{def}}{=} & \mathsf{A}(\mathsf{Happens}\ \alpha) \\ (\mathsf{Doesn't}\ \alpha) & \stackrel{\text{def}}{=} & \mathsf{A}\neg(\mathsf{Happens}\ \alpha)\end{array}$$

Finally, we find it convenient to make use of *mutual* mental states, although we recognise that such states are idealisations, not realisable in any system that admits the possibility of communication failure [9]. The mutual belief of φ in a group of agents g is written $(\mathsf{M\text{-}Bel}\ g\ \varphi)$, and the mutual goal of φ in g is written $(\mathsf{M\text{-}Goal}\ g\ \varphi)$. We give the full definition of mutual belief, but omit that for M-Goal, since it is essentially identical. Mutual belief is defined via an 'everyone believes' operator, E-Bel, which plays the role of the 'everyone knows' operator in knowledge theory (see, e.g., [9, p23]).

$$\begin{array}{rcl}(\mathsf{E\text{-}Bel}\ g\ \varphi\ 0) & \stackrel{\text{def}}{=} & \varphi \\ (\mathsf{E\text{-}Bel}\ g\ \varphi\ u+1) & \stackrel{\text{def}}{=} & \forall i \cdot (i \in g) \Rightarrow (\mathsf{Bel}\ i\ (\mathsf{E\text{-}Bel}\ g\ \varphi\ u)) \\ (\mathsf{M\text{-}Bel}\ g\ \varphi) & \stackrel{\text{def}}{=} & (\mathsf{E\text{-}Bel}\ g\ \varphi\ u) \quad \text{for all } u \in \mathbb{N}\end{array}$$

4 Commitments, Conventions, and Intentions

The key mental states that control agent behaviour in our model are intentions and joint intentions — the former define local asocial behaviour, the latter control social behaviour. Intentions are so central because they provide both the stability and predictability that is necessary for social interaction, and the flexibility and reactivity that is necessary to cope with a changing environment. Previous attempts to formalise commitment have not distinguished between the *commitment* that underpins

an intention and the associated *convention*. We clearly distinguish the two concepts: a *commitment* is a pledge or a promise; a *convention* is a means of monitoring a commitment — it specifies under what circumstances a commitment can be abandoned and how an agent should behave both locally and towards others when one of these conditions arises [16].

In more detail, one may commit either to a particular course of action, or, more generally, to a state of affairs. Here, we are concerned only with commitments that are *future-directed* towards a state of affairs. Commitments have a number of important properties (see [16] and [5, pp217–219] for a discussion), but the most important is that *commitments persist*: having adopted a commitment, we do not expect an agent to drop it until, for some reason, it becomes redundant. The conditions under which a commitment can become redundant are specified in the associated convention — examples include the motivation for the goal no longer being present, the goal being achieved, and the realisation that the goal will never be achieved [5].

When a group of agents are engaged in a cooperative activity they must have a joint commitment to the overall aim, as well as their individual commitments to the specific tasks that they have been assigned. This joint commitment shares the persistence property of the individual commitment; however it differs in that its state is distributed amongst the team members. To minimise the potential drawbacks of this distribution, an appropriate social convention must be put in place. This social convention identifies the conditions under which the joint commitment can be dropped, and also describes how the agent should behave towards its fellow team members. For example, if an agent drops its joint commitment because it believes that the goal will never be attained, then it is part of the notion of 'cooperativeness' which is inherent in joint action that it informs all of its fellow team members of its change of state. In this context, social conventions provide general guidelines, and a common frame of reference in which agents can work. By adopting a convention, every agent knows what is expected both of it, and of every other agent, as part of the collective working towards the goal, and knows that every other agent has a similar set of expectations.

Having informally introduced commitments and conventions, we now present rigorous definitions. A convention is a set of rules, each rule consisting of a re-evaluation condition ρ and a goal γ. The idea is that if ever an agent believes ρ to be true, then it must adopt γ as a goal, and keep this goal until the commitment becomes redundant. Formally, a *convention*, c, is an indexed set of pairs:

$$c = \{(\rho_k, \gamma_k) \mid k \in \{1, \ldots, l\}\}$$

where ρ_k is a *re-evaluation condition*, and γ_k is a *goal*, for all $1 \leq k \leq l$.

Joint commitments have a number of parameters. First, a joint commitment is held by a group g of agents. Second, joint commitments are held with respect to some goal φ; this is the state of affairs that the group is committed to bringing about. Third, joint commitments are held relative to a *motivation*, which characterises the justification for the commitment. They also have a *pre-condition*, which describes what must initially be true of the world in order for the commitment to be held. For example, in most types of joint commitment, we do not expect participating agents to initially believe that the object of the commitment, φ, is true. Finally, a joint commitment is parameterised by a convention c. Joint commitment is then

informally defined as follows. A group g is jointly committed to a goal φ with respect to motivation ψ, pre-condition χ, and convention c iff:

1. pre-condition χ is initially satisfied; and
2. every agent $i \in g$ has a goal of φ until the termination condition is satisfied;
3. until the termination condition is satisfied, if any agent $i \in g$ believes that the re-evaluation condition of any rule in c is satisfied, then it adopts the goal corresponding to the re-evaluation condition, and maintains this goal until the termination condition is satisfied.

where the termination condition is that one of the goal parts of the convention rules is satisfied. More formally, if $c = \{(\rho_k, \gamma_k) \mid k \in \{1, \ldots, l\}\}$ is a convention, then:

$$(\mathsf{J\text{-}Commit}\ g\ \varphi\ \psi\ \chi\ c) \stackrel{\text{def}}{=} \forall i \cdot (i \in g) \Rightarrow \chi \wedge \mathsf{A}((p \wedge q)\,\mathcal{U}\,r)$$

where

$$p \stackrel{\text{def}}{=} (\mathsf{Goal}\ i\ \varphi)$$

and

$$q \stackrel{\text{def}}{=} \bigwedge_{k=1}^{l} (\mathsf{Bel}\ i\ \rho_k) \Rightarrow \mathsf{A}[(\mathsf{Goal}\ i\ \gamma_k)\,\mathcal{U}\,r]$$

and

$$r \stackrel{\text{def}}{=} \bigvee_{k=1}^{l} \gamma_k.$$

Notice that the motivation, ψ does not appear to be used in the right hand side of this definition; however, it can appear in the convention rules. To illustrate how commitments and conventions work, we will specify a *minimal social convention*, that is similar to the Levesque-Cohen model of joint persistent goals (JPGs) [18]. Let

$$\chi_{soc} \stackrel{\text{def}}{=} \neg(\mathsf{Bel}\ i\ \varphi) \wedge (\mathsf{Bel}\ i\ \mathsf{E}\Diamond\varphi)$$

and

$$c_{soc} \stackrel{\text{def}}{=} \left\{ \begin{array}{l} (\underbrace{(\mathsf{Bel}\ i\ \varphi)}_{\rho_1}, \underbrace{(\mathsf{M\text{-}Bel}\ g\ \varphi)}_{\gamma_1}), \\ (\underbrace{(\mathsf{Bel}\ i\ \mathsf{A}\,\Box\neg\varphi)}_{\rho_2}, \underbrace{(\mathsf{M\text{-}Bel}\ g\ \mathsf{A}\,\Box\neg\varphi)}_{\gamma_2}), \\ (\underbrace{(\mathsf{Bel}\ i\ \neg\psi)}_{\rho_3}, \underbrace{(\mathsf{M\text{-}Bel}\ g\ \neg\psi)}_{\gamma_3}) \end{array} \right\}.$$

It is not difficult to see that $(\mathsf{J\text{-}Commit}\ g\ \varphi\ \psi\ \chi_{soc}\ c_{soc})$ expands to:

$$
\begin{array}{l}
\forall i \cdot (i \in g) \Rightarrow \\
\neg(\mathsf{Bel}\ i\ \varphi) \wedge (\mathsf{Bel}\ i\ \mathsf{E}\Diamond\varphi) \wedge \\
\mathsf{A}\left[\left(\begin{array}{lcll}
(\mathsf{Goal}\ i\ \varphi) \wedge & & & \\
(\ ((\mathsf{Bel}\ i\ \varphi) & \Rightarrow & \mathsf{A}((\mathsf{Goal}\ i\ (\mathsf{M\text{-}Bel}\ g\ \varphi))\,\mathcal{U}\,p)) & \wedge \\
\quad ((\mathsf{Bel}\ i\ \mathsf{A}\,\Box\neg\varphi) & \Rightarrow & \mathsf{A}((\mathsf{Goal}\ i\ (\mathsf{M\text{-}Bel}\ g\ \mathsf{A}\,\Box\neg\varphi))\,\mathcal{U}\,p)) & \wedge \\
\quad ((\mathsf{Bel}\ i\ \neg\psi) & \Rightarrow & \mathsf{A}((\mathsf{Goal}\ i\ (\mathsf{M\text{-}Bel}\ g\ \neg\psi))\,\mathcal{U}\,p)) & \\
\ )
\end{array}\right)\ \mathcal{U}\,p\right]
\end{array}
$$

where

$$p \stackrel{\text{def}}{=} [(\mathsf{M\text{-}Bel}\ g\ \varphi) \vee (\mathsf{M\text{-}Bel}\ g\ \mathsf{A}\,\Box\neg\varphi) \vee (\mathsf{M\text{-}Bel}\ g\ \neg\psi)].$$

A collective with a such a commitment will have a mental state in which:

- initially, every agent does not believe that the goal φ is satisfied, but believes φ is possible;
- every agent i then has a goal of φ until the termination condition is satisfied (see below);
- until the termination condition is satisfied, then:
  - if any agent i believes that the goal is achieved, then it will have a goal that this becomes a mutual belief, and will retain this goal until the termination condition is satisfied;
  - if any agent i believes that the goal is impossible, then it will have a goal that this becomes a mutual belief, and will retain this goal until the termination condition is satisfied;
  - if any agent i believes that the motivation ψ for the goal is no longer present, then it will have a goal that this becomes a mutual belief, and will retain this goal until the termination condition is satisfied;
- the termination condition is that it is mutually believed that either:
  - the goal φ is satisfied;
  - the goal φ is impossible to achieve;
  - the motivation/justification ψ for the goal is no longer present.

To represent systems in which commitments can be dropped for different reasons, then all that needs to be changed is the convention. This flexibility is only available because conventions are clearly identified as a separate concept — for example in [18] the above conditions are hardwired into the definition of joint commitment, hence the model builders are imposing a definitive convention and there is no scope for varying agent behaviour according to the complexity of the collaboration.

We use the model of joint commitments to define joint intentions, which are held by a group g with respect to an action α and motivation ψ.

$$
\begin{array}{l}
(\mathsf{J\text{-}Intend}\ g\ \alpha\ \psi) \stackrel{\text{def}}{=}\ (\mathsf{M\text{-}Bel}\ g\ (\mathsf{Agts}\ \alpha\ g)) \wedge \\
\qquad\qquad (\mathsf{J\text{-}Commit}\ \ g\ \mathsf{A}\Diamond(\mathsf{Happens}\ (\mathsf{M\text{-}Bel}\ g\ (\mathsf{Does}\ \alpha))?;\alpha)\ \ \psi\ \chi_{soc}\ c_{soc})
\end{array}
$$

This definition is based on that by Levesque-Cohen [18, p98]: the idea is that having a joint intention to do α means having a joint commitment that eventually g will believe α will happen next, and then α happens next. Note that we could make commitments and conventions a parameter of joint intentions; we do not do this in order to simplify subsequent formalism.

We define individual commitments as a special case of joint commitment.

$$(\mathsf{Commit}\ i\ \varphi\ \psi) \stackrel{\text{def}}{=} \forall g \cdot (\mathsf{Singleton}\ g\ i) \Rightarrow (\mathsf{J\text{-}Commit}\ g\ \varphi\ \psi\ \chi_{soc}\ c_{soc})$$

An individual intention by agent i to do α with respect to motivation ψ is similarly defined as a special case of joint intention.

$$(\mathsf{Intend}\ i\ \alpha\ \psi) \stackrel{\text{def}}{=} \forall g \cdot (\mathsf{Singleton}\ g\ i) \Rightarrow (\mathsf{J\text{-}Intend}\ g\ \alpha\ \psi)$$

5 The Cooperative Problem Solving Process

In this section, we present the main contribution of the paper: a four-stage model of CPS. First, an overview is given. Each stage in the model is subsequently considered in more detail, and then formalised. The four stages are:

1. *Recognition*: The CPS process begins when some agent recognises the potential for cooperative action. This recognition may come about because an agent has a goal that it does not have the ability to achieve on its own, or else because the agent prefers a cooperative solution.

2. *Team formation*: During this stage, the agent that recognised the potential for cooperative action at stage (1) solicits assistance. If this stage is successful, then it will end with a group of agents having some kind of nominal commitment to collective action.

3. *Plan formation*: During this stage, the agents attempt to negotiate a joint plan which they believe will achieve the desired goal.

4. *Team action*: During this stage, the newly agreed plan of joint action is executed by the agents, which maintain a close-knit relationship throughout. This relationship is defined by a *convention*, which every agent follows.

Although we believe that most instances of CPS exhibit these stages in some form (either explicitly or implicitly), we stress that the model is *idealised*. We recognise that there are cases which the model cannot account for, and we have attempted to highlight such cases wherever appropriate. Our aim is to construct a framework that is complete, (in that it describes CPS from beginning to end), but abstract, (in that details which might obscure more significant points have been omitted). Finally, we note that in reality, these four stages are *iterative*, in that if one stage fails, the agents may return to previous stages. In the interests of simplicity, we have not attempted to represent this aspect in our model.

5.1 Recognition

CPS begins when some agent in a multi-agent community has a goal, and recognises the potential for cooperative action with respect to that goal. Recognition may occur for several reasons. The paradigm case is that in which the agent is unable to achieve the goal in isolation, but believes that cooperative action can achieve it. For example, an agent may have a goal which, to achieve, requires information that is only accessible to another agent. Without the cooperation of this other agent, the goal cannot be achieved. More prosaically, an agent with a goal to move a heavy object might simply not have the strength to do this alone.

Alternatively, an agent may be able to achieve the goal on its own, but may not want to. There may be several reasons for this. First, it may believe that in working alone, it will clobber one of its other goals. For example, suppose I have a goal of lifting a heavy object. I may have the capability of lifting the object, but I might believe that in so doing, I would injure my back, thereby clobbering my goal of being healthy. In this case, a cooperative solution — involving no injury to my back — is preferable. More generally, an agent may believe that a cooperative solution will in some way be better than a solution achieved by action in isolation. For example, a solution might be obtained more quickly, or may be more accurate as a result of cooperative action.

Believing that you either cannot achieve your goal in isolation, or that, (for whatever reason), you would prefer not to work alone, is part of the potential for cooperation. But it is not enough in itself to initiate the social process. For there to be potential for cooperation with respect to an agent's goal, the agent must also believe there is some group of agents that can actually achieve the goal.

In order to precisely define the conditions that characterise the potential for cooperative action, it is necessary to introduce a number of subsidiary definitions. First, we require definitions of single-agent and multi-agent *ability*: what it means to be *able* to bring about some state of the world. Rather than complicate the logic further by introducing yet another primitive modality, we adapt a well-known definition of ability that was originally proposed by Moore [19].

As a first attempt to define ability, we might say an agent has the ability to achieve some state φ if it knows of an action that it can perform, which would be guaranteed to achieve the state of affairs. We will call this *type 1 ability*, and define it as follows.

$$(\mathsf{Able}_1\ i\ \varphi) \stackrel{\text{def}}{=} \exists \alpha \cdot (\mathsf{Bel}\ i\ (\mathsf{Agt}\ \alpha\ i) \wedge (\mathsf{Achieves}\ \alpha\ \varphi)) \wedge (\mathsf{Agt}\ \alpha\ i) \wedge (\mathsf{Achieves}\ \alpha\ \varphi)$$

Note that the action α in this definition is quantified *de re* with respect to the Bel modality [15, p183]. The significance of this is that the agent must be 'aware of the identity' of the action — it must have a *rigid designator* for it. Thus it is not enough for the agent to believe that there exists *some* action that will achieve the goal. It must be aware of exactly *which* action will achieve it.

Before proceeding, we prove some results about type 1 ability. First, we show that if an agent has the type 1 ability to bring about some state of affairs, then that state of affairs is actually possible.

Theorem 1 $\models (\mathsf{Able}_1\ i\ \varphi) \Rightarrow \mathsf{E}\Diamond\varphi$

Proof Assume that $\langle M, V, s\rangle \models (\mathsf{Able}_1\ i\ \varphi)$ for arbitrary $\langle M, V, s\rangle$. By expanding out the definition of Able_1, we get $\langle M, V, s\rangle \models \exists\alpha \cdot (\mathsf{Achieves}\ \alpha\ \varphi)$. From this and the definition of Achieves we get $\langle M, V, s\rangle \models \mathsf{E}(\mathsf{Happens}\ \alpha) \wedge \mathsf{A}(\mathsf{Happens}\ \alpha; \varphi?)$, and hence $\langle M, V, s\rangle \models \mathsf{E}\Diamond\varphi$.

If an agent has the type 1 ability to bring about a state of affairs, then it is aware of this.

Theorem 2 $\models (\mathsf{Able}_1\ i\ \varphi) \Rightarrow (\mathsf{Bel}\ i\ (\mathsf{Able}_1\ i\ \varphi))$

Proof We need to show that if $\langle M, V, s\rangle \models (\mathsf{Able}_1\ i\ \varphi)$ for arbitrary $\langle M, V, s\rangle$, then $\langle M, V, s'\rangle \models (\mathsf{Able}_1\ i\ \varphi)$ for all $s' \in S$ such that $(s, s') \in B([\![i]\!])$. Start by assuming that $\langle M, V, s\rangle \models (\mathsf{Able}_1\ i\ \varphi)$ for arbitrary $\langle M, V, s\rangle$. Hence $\langle M, V, s\rangle \models \exists\alpha \cdot (\mathsf{Bel}\ i\ (\mathsf{Agt}\ \alpha\ i) \wedge (\mathsf{Achieves}\ \alpha\ \varphi))$, and so for all $s' \in S$ such that $(s, s') \in B([\![i]\!])$, we have $\langle M, V, s'\rangle \models \exists\alpha \cdot (\mathsf{Agt}\ \alpha\ i) \wedge (\mathsf{Achieves}\ \alpha\ \varphi)$. We want to show that $\langle M, V, s''\rangle \models (\mathsf{Agt}\ \alpha\ i) \wedge (\mathsf{Achieves}\ \alpha\ \varphi)$ for all $s'' \in S$ such that $(s', s'') \in B([\![i]\!])$. But since the belief accessibility relation B is transitive, it must be that if $(s, s') \in B([\![i]\!])$, and $(s', s'') \in B([\![i]\!])$, then $(s, s'') \in B([\![i]\!])$. Hence $\langle M, V, s''\rangle \models (\mathsf{Agt}\ \alpha\ i) \wedge (\mathsf{Achieves}\ \alpha\ \varphi)$, so $\langle M, V, s'\rangle \models \exists\alpha \cdot (\mathsf{Bel}\ i\ (\mathsf{Agt}\ \alpha\ i) \wedge (\mathsf{Achieves}\ \alpha\ \varphi)) \wedge (\mathsf{Agt}\ \alpha\ i) \wedge (\mathsf{Achieves}\ \alpha\ \varphi)$, and hence $\langle M, V, s'\rangle \models (\mathsf{Able}_1\ i\ \varphi)$, and we are done.

If an agent has the type 1 ability to bring about some state of affairs, then it believes that state of affairs is possible.

Theorem 3 $\models (\mathsf{Able}_1\ i\ \varphi) \Rightarrow (\mathsf{Bel}\ i\ \mathsf{E}\Diamond\varphi)$

Proof Assume that $\langle M, V, s\rangle \models (\mathsf{Able}_1\ i\ \varphi)$ for arbitrary $\langle M, V, s\rangle$. By Theorem 2, we therefore have $\langle M, V, s'\rangle \models (\mathsf{Able}_1\ i\ \varphi)$ for all $s' \in S$ such that $(s, s') \in B([\![i]\!])$. From Theorem 1, we thus have $\langle M, V, s'\rangle \models \mathsf{E}\Diamond\varphi$, and we are done.

An obvious failing of this definition when measured against our intuitions about ability is that it fails to allow for an agent performing an action in order to *find out* how to bring about some state of affairs. This motivates a definition of *type 2 ability*, which allows for the possibility of an agent performing an action in order to find out how to bring about a state of affairs. The idea is that an agent will have the ability to bring about a state of affairs φ if either it has the type 1 ability to bring about φ (i.e., it knows of some action that it could perform, which is guaranteed to bring about φ directly), or else it has the type 1 ability to bring about a state of affairs where it has the type 1 ability to bring about φ. We formalise this as follows.

$$(\mathsf{Able}\ i\ \varphi) \stackrel{\text{def}}{=} (\mathsf{Able}_1\ i\ \varphi) \vee (\mathsf{Able}_1\ i\ (\mathsf{Able}_1\ i\ \varphi))$$

It is straightforward to see that type 1 ability implies type 2 ability.

Theorem 4 $\models (\mathsf{Able}_1\ i\ \varphi) \Rightarrow (\mathsf{Able}\ i\ \varphi)$

We can also prove results analogous to Theorems 1, 2, and 3 for Able. (Proofs for Theorems 5 and 6 are straightforward, and are therefore omitted.)

Theorem 5 $\models (\mathsf{Able}\ i\ \varphi) \Rightarrow \mathsf{E}\Diamond\varphi$

Theorem 6 $\models (\mathsf{Able}\ i\ \varphi) \Rightarrow (\mathsf{Bel}\ i\ \mathsf{E}\Diamond\varphi)$

Theorem 7 $\models (\mathsf{Able}\ i\ \varphi) \Rightarrow (\mathsf{Bel}\ i\ (\mathsf{Able}\ i\ \varphi))$

Proof We need to show that if $\langle M, V, s\rangle \models (\mathsf{Able}\ i\ \varphi)$ for arbitrary $\langle M, V, s\rangle$, then $\langle M, V, s'\rangle \models (\mathsf{Able}\ i\ \varphi)$ for all $s' \in S$ such that $(s, s') \in B([\![i]\!])$. Start by assuming that $\langle M, V, s\rangle \models (\mathsf{Able}\ i\ \varphi)$ for arbitrary $\langle M, V, s\rangle$. By expanding out the definition of Able, we get $\langle M, V, s\rangle \models (\mathsf{Able}_1\ i\ \varphi) \vee (\mathsf{Able}_1\ i\ (\mathsf{Able}_1\ i\ \varphi))$. We thus reason by cases. In the first case, we have $\langle M, V, s\rangle \models (\mathsf{Able}_1\ i\ \varphi)$, and so by Theorem 2, we have $\langle M, V, s\rangle \models (\mathsf{Bel}\ i\ (\mathsf{Able}_1\ i\ \varphi))$, hence $\langle M, V, s'\rangle \models (\mathsf{Able}_1\ i\ \varphi)$ for all $s' \in S$ such that $(s, s') \in B([\![i]\!])$. By Theorem 4, we therefore have $\langle M, V, s'\rangle \models (\mathsf{Able}\ i\ \varphi)$ for all $s' \in S$ such that $(s, s') \in B([\![i]\!])$, and so $\langle M, V, s\rangle \models (\mathsf{Bel}\ i\ (\mathsf{Able}\ i\ \varphi))$. In the second case, we have $\langle M, V, s\rangle \models (\mathsf{Able}_1\ i\ (\mathsf{Able}_1\ i\ \varphi))$. So by Theorem 2, we have $\langle M, V, s\rangle \models (\mathsf{Bel}\ i\ (\mathsf{Able}_1\ i\ (\mathsf{Able}_1\ i\ \varphi)))$, hence $\langle M, V, s'\rangle \models (\mathsf{Able}_1\ i\ (\mathsf{Able}_1\ i\ \varphi))$ for all $s' \in S$ such that $(s, s') \in B([\![i]\!])$, thus $\langle M, V, s'\rangle \models (\mathsf{Able}\ i\ \varphi)$ for all $s' \in S$ such that $(s, s') \in B([\![i]\!])$, and we are done.

To simplify future definitions, we will introduce another derived operator, Unable, which has the obvious interpretation.

$$(\mathsf{Unable}\ i\ \varphi) \stackrel{\text{def}}{=} \neg(\mathsf{Able}\ i\ \varphi)$$

We shall assume that if an agent is unable to achieve some state of affairs, then it is aware that it is unable to achieve this.

$$\models (\mathsf{Unable}\ i\ \varphi) \Rightarrow (\mathsf{Bel}\ i\ (\mathsf{Unable}\ i\ \varphi)) \tag{1}$$

We now need to define *multi-agent ability*, which we do by simply adapting the definition of single-agent ability to the multi-agent case.

$$\begin{aligned}(\mathsf{J\text{-}Able}_1\ g\ \varphi) &\stackrel{\text{def}}{=} \exists\alpha \cdot (\mathsf{M\text{-}Bel}\ g\ (\mathsf{Agts}\ \alpha\ g) \wedge (\mathsf{Achieves}\ \alpha\ \varphi)) \wedge \\ &\qquad (\mathsf{Agts}\ \alpha\ g) \wedge (\mathsf{Achieves}\ \alpha\ \varphi) \\ (\mathsf{J\text{-}Able}\ g\ \varphi) &\stackrel{\text{def}}{=} (\mathsf{J\text{-}Able}_1\ g\ \varphi) \vee (\mathsf{J\text{-}Able}_1\ g\ (\mathsf{J\text{-}Able}_1\ g\ \varphi))\end{aligned}$$

We now present some results about joint ability that are analogous to Theorems 1 through to 7. (We omit proofs that can be obtained by straightforward adaptations of earlier results.)

Theorem 8 $\models (\mathsf{J\text{-}Able}_1\ g\ \varphi) \Rightarrow \mathsf{E}\Diamond\varphi$

Theorem 9 $\models (\mathsf{J\text{-}Able}_1\ g\ \varphi) \Rightarrow (\mathsf{M\text{-}Bel}\ g\ \mathsf{E}\Diamond\varphi)$

Theorem 10 $\models (\mathsf{J\text{-}Able}_1\ g\ \varphi) \Rightarrow (\mathsf{M\text{-}Bel}\ g\ (\mathsf{J\text{-}Able}_1\ g\ \varphi))$

Theorem 11 $\models (\mathsf{J\text{-}Able}_1\ g\ \varphi) \Rightarrow (\mathsf{J\text{-}Able}\ g\ \varphi)$

Theorem 12 $\models (\mathsf{J\text{-}Able}\ g\ \varphi) \Rightarrow \mathsf{E}\Diamond\varphi$

Theorem 13 $\models (\mathsf{J\text{-}Able}\ g\ \varphi) \Rightarrow (\mathsf{M\text{-}Bel}\ g\ (\mathsf{J\text{-}Able}\ g\ \varphi))$

Proof We need to show that if $\langle M, V, s\rangle \models (\mathsf{J\text{-}Able}\ g\ \varphi)$ for arbitrary $\langle M, V, s\rangle$, then $\langle M, V, s\rangle \models (\mathsf{M\text{-}Bel}\ g\ (\mathsf{J\text{-}Able}\ g\ \varphi))$. Start by assuming that $\langle M, V, s\rangle \models (\mathsf{J\text{-}Able}\ g\ \varphi)$ for arbitrary $\langle M, V, s\rangle$. By expanding out the definition of J-Able, we get $\langle M, V, s\rangle \models (\mathsf{J\text{-}Able}_1\ g\ \varphi) \vee (\mathsf{J\text{-}Able}_1\ g\ (\mathsf{J\text{-}Able}_1\ g\ \varphi))$. We thus reason by cases. In the first case, we have $\langle M, V, s\rangle \models (\mathsf{J\text{-}Able}_1\ g\ \varphi)$, and so by Theorem 10, we have $\langle M, V, s\rangle \models (\mathsf{M\text{-}Bel}\ g\ (\mathsf{J\text{-}Able}_1\ g\ \varphi))$, so from Theorem 11, we have $\langle M, V, s\rangle \models (\mathsf{M\text{-}Bel}\ g\ (\mathsf{J\text{-}Able}\ g\ \varphi))$. In the second case, we have $\langle M, V, s\rangle \models (\mathsf{J\text{-}Able}_1\ g\ (\mathsf{J\text{-}Able}_1\ g\ \varphi))$. So by Theorem 10, we have $\langle M, V, s\rangle \models (\mathsf{M\text{-}Bel}\ g\ (\mathsf{J\text{-}Able}_1\ g\ (\mathsf{J\text{-}Able}_1\ g\ \varphi)))$, thus $\langle M, V, s\rangle \models (\mathsf{M\text{-}Bel}\ g\ (\mathsf{J\text{-}Able}\ g\ \varphi))$, and we are done.

Theorem 14 $\models (\mathsf{J\text{-}Able}\ g\ \varphi) \Rightarrow (\mathsf{M\text{-}Bel}\ i\ \mathsf{E}\Diamond\varphi)$

Proof Straightforward from Theorems 13 and 12.

We can now more precisely define potential for cooperation. With respect to agent i's goal φ, there is potential for cooperation iff:

1. there is some group g such that i believes that g can jointly achieve φ;

and either

2. i can't achieve φ in isolation; or
3. i believes that for every action α that it could perform that achieves φ, it has a goal of not performing α.

Note that in clause (1), an agent needs to know the identity of the group that it believes can cooperate to achieve its goal. This is perhaps an over-strong assumption. It precludes an agent attempting to find out the identity of a group that can achieve the goal, and it does not allow an agent to simply broadcast its goal in the hope of attracting help (as in the Contract Net protocol [27]). We leave such refinements to future work. Clause (2) represents the paradigm reason for an agent considering a cooperative solution: because it is unable to achieve the goal on its own. Clause (3) defines the alternative reason for an agent considering cooperation: it prefers not to perform any of the actions that might achieve the goal. (We do not consider the reasons why an agent will not want to perform a particular action — this will be domain-specific.)

Using the various definitions above, we can now formally state the conditions that characterise the potential for cooperation.

$$
\begin{aligned}
(\mathsf{PfC}\ i\ \varphi) \stackrel{\text{def}}{=}\ & (\mathsf{Goal}\ i\ \varphi) \wedge \\
& (\mathsf{Bel}\ i\ \neg\varphi) \wedge \\
& \exists g \cdot (\mathsf{Bel}\ i\ (\mathsf{J\text{-}Able}\ g\ \varphi)) \wedge \\
& \left[\begin{array}{l} (\mathsf{Unable}\ i\ \varphi) \vee \\ (\mathsf{Bel}\ i\ \forall\alpha \cdot (\mathsf{Agt}\ \alpha\ i) \wedge (\mathsf{Achieves}\ \alpha\ \varphi) \Rightarrow (\mathsf{Goal}\ i\ (\mathsf{Doesn't}\ \alpha))) \end{array}\right]
\end{aligned}
$$

We now prove some properties of potential for cooperation.

Theorem 15 $\models (\mathsf{PfC}\ i\ \varphi) \Rightarrow (\mathsf{Bel}\ i\ \mathsf{E}\Diamond\varphi)$

Proof Assume $\langle M, V, s\rangle \models (\mathsf{PfC}\ i\ \varphi)$ for arbitrary $\langle M, V, s\rangle$. Then by expanding out the definition of PfC, we get $\langle M, V, s\rangle \models \exists g \cdot (\mathsf{Bel}\ i\ (\mathsf{J\text{-}Able}\ g\ \varphi))$. So for all $s' \in S$ such that $(s, s') \in B([\![i]\!])$, we have $\langle M, V, s'\rangle \models (\mathsf{J\text{-}Able}\ g\ \varphi)$, and so from Theorem 12, we have $\langle M, V, s'\rangle \models \mathsf{E}\Diamond\varphi$. Thus $\langle M, V, s\rangle \models (\mathsf{Bel}\ i\ \mathsf{E}\Diamond\varphi)$.

The final result of this section shows that if there is potential for cooperation with respect to an agent's goal, then the agent is aware of this.

Theorem 16 $\models (\mathsf{PfC}\ i\ \varphi) \Rightarrow (\mathsf{Bel}\ i\ (\mathsf{PfC}\ i\ \varphi))$

Proof Assume $\langle M, V, s\rangle \models (\mathsf{PfC}\ i\ \varphi)$, for arbitrary $\langle M, V, s\rangle$. We need to show that $\langle M, V, s\rangle \models (\mathsf{Bel}\ i\ \chi)$, for each conjunct χ in the definition of potential for cooperation:

- $\langle M, V, s\rangle \models (\mathsf{Bel}\ i\ (\mathsf{Goal}\ i\ \varphi))$

  Immediate from (16).

- $\langle M, V, s\rangle \models (\mathsf{Bel}\ i\ (\mathsf{Bel}\ i\ \neg\varphi))$

  Immediate from axiom 4 for belief modalities.

- $\langle M, V, s\rangle \models (\mathsf{Bel}\ i\ \exists g \cdot (\mathsf{Bel}\ i\ (\mathsf{J\text{-}Able}\ g\ \varphi)))$

  Immediate from axiom 4 for belief modalities.

- $\langle M, V, s\rangle \models (\mathsf{Bel}\ i\ (\mathsf{Unable}\ i\ \varphi) \vee (\mathsf{Bel}\ i\ \forall\alpha \cdot (\mathsf{Agt}\ \alpha\ i) \wedge (\mathsf{Achieves}\ \alpha\ \varphi) \Rightarrow (\mathsf{Goal}\ i\ (\mathsf{Doesn't}\ \alpha))))$

  There are two cases to consider. For the first case, assume that $\langle M, V, s\rangle \models (\mathsf{Unable}\ i\ \varphi)$. Then by the assumption that agents are aware of what they cannot achieve, $\langle M, V, s\rangle \models (\mathsf{Bel}\ i\ (\mathsf{Unable}\ i\ \varphi))$.

  For the second case, assume $\langle M, V, s\rangle \models (\mathsf{Bel}\ i\ \forall\alpha \cdot (\mathsf{Agt}\ \alpha\ i) \wedge (\mathsf{Achieves}\ \alpha\ \varphi) \Rightarrow (\mathsf{Goal}\ i\ (\mathsf{Doesn't}\ \alpha)))$. In this case, $\langle M, V, s\rangle \models (\mathsf{Bel}\ i\ (\mathsf{Bel}\ i\ \forall\alpha \cdot (\mathsf{Agt}\ \alpha\ i) \wedge (\mathsf{Achieves}\ \alpha\ \varphi) \Rightarrow (\mathsf{Goal}\ i\ (\mathsf{Doesn't}\ \alpha))))$ follows from axiom 4 for belief modalities.

5.2 Team Formation

Having identified the potential for cooperative action with respect to one of its goals, what is a rational agent to do? We propose that such an agent will attempt to *solicit assistance* from a group of agents that it believes can achieve the goal. If the agent is successful, then at the conclusion of this *team formation* stage, the agent will have brought about in such a group a mental state wherein each member of the group has a nominal commitment to collective action. The group will not yet have fixed upon an action to perform, and in fact will not share any kind of commitment other than to the *principle* of joint action. In particular, there will not yet be a joint intention: this comes later.

How does an agent go about forming a team? The most important point to note is that it cannot *guarantee* that it will be successful in forming a team: it can only *attempt* it. We therefore require a model of attempts. We adopt that proposed by Cohen and Levesque [6, p240]. The idea is that an attempt by agent i to bring about a state φ is an action α, which is performed by i with the goal that after α is performed, φ is satisfied, but with the intention that at least ψ is satisfied. The ultimate goal of

the attempt — the thing that i hopes to bring about — is represented by φ, whereas ψ represents 'what it takes to make an honest effort' [6, p240]. If i is successful, then bringing about ψ will be sufficient to cause φ.

Formally, an attempt by i to achieve φ by performing α, at least achieving ψ, is written $\{\mathsf{Attempt}\ i\ \alpha\ \varphi\ \psi\}$; following Cohen and Levesque, we use curly brackets here to indicate that attempts are complex actions, rather than predicates or modal operators [6, p240].

$$\{\mathsf{Attempt}\ i\ \alpha\ \varphi\ \psi\} \stackrel{\text{def}}{=} \left[\begin{array}{l} (\mathsf{Bel}\ i\ \neg\varphi) \wedge \\ (\mathsf{Agt}\ \alpha\ i) \wedge \\ (\mathsf{Goal}\ i\ (\mathsf{Achieves}\ \alpha\ \varphi)) \wedge \\ (\mathsf{Intend}\ i\ \alpha; \psi?\ \mathsf{true}) \end{array}\right] ?; \alpha$$

We prove some properties of attempts. First, if an agent attempts to bring about some state of affairs φ, then it believes that φ is possible.

Theorem 17 $\models (\mathsf{Happens}\ \{\mathsf{Attempt}\ i\ \alpha\ \varphi\ \psi\}) \Rightarrow (\mathsf{Bel}\ i\ \mathsf{E}\Diamond\varphi)$

Proof Assume $\langle M, V, p\rangle \models (\mathsf{Happens}\ \{\mathsf{Attempt}\ i\ \alpha\ \varphi\ \psi\})$ for arbitrary $\langle M, V, p\rangle$. Then from the definition of Attempt, we have $\langle M, V, p\rangle \models (\mathsf{Goal}\ i\ (\mathsf{Achieves}\ \alpha\ \varphi))$. Hence from the properties of Achieves, we have $\langle M, V, p\rangle \models (\mathsf{Goal}\ i\ \mathsf{E}\Diamond\varphi)$. Now from (17), it must be that $\langle M, V, p\rangle \models (\mathsf{Bel}\ i\ \mathsf{E}\Diamond\varphi)$.

Similarly, if an agent attempts to bring about some state of affairs φ by at least bringing about ψ, then it believes that ψ is possible.

Theorem 18 $\models (\mathsf{Happens}\ \{\mathsf{Attempt}\ i\ \alpha\ \varphi\ \psi\}) \Rightarrow (\mathsf{Bel}\ i\ \mathsf{E}\Diamond\psi)$

Proof Assume $\langle M, V, p\rangle \models (\mathsf{Happens}\ \{\mathsf{Attempt}\ i\ \alpha\ \varphi\ \psi\})$ for arbitrary $\langle M, V, p\rangle$. Then from the definition of Attempt, we have we have $\langle M, V, p\rangle \models (\mathsf{Intend}\ i\ \alpha; \psi?\ \mathsf{true})$. Now, from the properties of Intend, we can conclude that $\langle M, V, p\rangle \models (\mathsf{Bel}\ i\ \mathsf{E}\Diamond(\mathsf{Happens}\ \alpha; \psi?))$. By temporal reasoning, we can conclude $\langle M, V, p\rangle \models (\mathsf{Bel}\ i\ \mathsf{E}\Diamond\psi)$.

The team formation stage can then be characterised as the following assumption about rational agents: an agent i, who believes that there is potential for cooperative action with respect to its goal φ, will eventually attempt to bring about in some group g, (which it believes can jointly achieve φ), a state wherein:

1. it is mutually believed in g that g can jointly achieve φ;
2. it is mutually believed in g that every agent in g is individually committed to φ, relative to i still having a goal of φ;

or, failing that, to at least cause in g

3. the mutual belief that i has a goal of φ; and
4. the mutual belief that i believes g can jointly achieve φ.

Parts (1) and (2) of this definition represent the minimal commitment that the group has towards i's goal φ if i is successful in its attempt to solicit assistance. This commitment does not yet involve a collective goal; merely the mutual belief that the group can bring about the goal, and that every member of the collective is individually committed to the goal on i's behalf. If g are helpfully inclined towards i, then this will be sufficient to cause them to proceed to the next stage of CPS. We leave as unspecified the reasons why g may (or may not) be helpfully inclined to i, as these reasons will be domain specific. Note that part (2) of the definition might arguably be dropped: an agent might have its own reasons for agreeing to participate in a cooperative action, that are unconnected with the original request for participation.

It is implicit within this assumption that agents are *veracious* with respect to their goals, i.e., that they will try to influence the group by revealing their true goal. We do not consider cases where agents are mendacious (i.e., they lie about their goals), or when agents do not reveal their goals. The interested reader is referred to [10, pp159–165] for a discussion and formalization of such considerations.

It is useful to introduce a definition which captures the commitment that agents have to collective action if team formation is successful. We write $(\mathsf{Pre\text{-}Team}\ g\ \varphi\ i)$ iff it is mutually believed in g that: (i) g can jointly achieve φ; and (ii) every agent in g has a commitment to φ, relative to i still having a goal of φ.

$$(\mathsf{Pre\text{-}Team}\ g\ \varphi\ i) \stackrel{\text{def}}{=} (\mathsf{M\text{-}Bel}\ g\ (\mathsf{J\text{-}Able}\ g\ \varphi) \wedge \forall j \cdot (j \in g) \Rightarrow (\mathsf{Commit}\ j\ \varphi\ (\mathsf{Goal}\ i\ \varphi)))$$

The following results capture some important properties of Pre-Team.

Theorem 19 $\models (\mathsf{Pre\text{-}Team}\ g\ \varphi\ i) \Rightarrow (\mathsf{M\text{-}Bel}\ g\ (\mathsf{Pre\text{-}Team}\ g\ \varphi\ i))$

Proof Straightforward from axiom 4 for M-Bel.

Theorem 20 $\models (\mathsf{Pre\text{-}Team}\ g\ \varphi\ i) \Rightarrow (\mathsf{M\text{-}Bel}\ g\ \mathsf{E}\Diamond\varphi)$

Proof Assume that $\langle M, V, s\rangle \models (\mathsf{Pre\text{-}Team}\ g\ \varphi\ i)$ for arbitrary $\langle M, V, s\rangle$. Expanding out the definition of Pre-Team gives $\langle M, V, s\rangle \models (\mathsf{M\text{-}Bel}\ g\ (\mathsf{J\text{-}Able}\ g\ \varphi))$. Theorem 12 tells us that $\models (\mathsf{J\text{-}Able}\ g\ \varphi) \Rightarrow \mathsf{E}\Diamond\varphi$, and from necessitation for M-Bel operators, we therefore know that $\models (\mathsf{M\text{-}Bel}\ g\ ((\mathsf{J\text{-}Able}\ g\ \varphi) \Rightarrow \mathsf{E}\Diamond\varphi))$. From the K axiom for M-Bel and propositional reasoning, we can therefore conclude that $\langle M, V, s\rangle \models (\mathsf{M\text{-}Bel}\ g\ \mathsf{E}\Diamond\varphi)$.

The main assumption concerning team formation can now be stated.

$$\models \forall i \cdot (\mathsf{Bel}\ i\ (\mathsf{PfC}\ i\ \varphi)) \Rightarrow \mathsf{A}\Diamond\exists g \cdot \exists \alpha \cdot (\mathsf{Happens}\ \{\mathsf{Attempt}\ i\ \alpha\ p\ q\}) \qquad (2)$$

where

$$p \stackrel{\text{def}}{=} (\mathsf{Pre\text{-}Team}\ g\ \varphi\ i)$$

and

$$q \stackrel{\text{def}}{=} (\mathsf{M\text{-}Bel}\ g\ (\mathsf{Goal}\ i\ \varphi) \wedge (\mathsf{Bel}\ i\ (\mathsf{J\text{-}Able}\ g\ \varphi))).$$

If team formation is successful, then for the first time there will be a *social* commitment: a commitment by a group of agents on behalf of another agent.

5.3 Plan Formation

If an agent is successful in its attempt to solicit assistance, then there will be a group of agents with a nominal commitment to collective action. But collective action cannot actually begin until the group agree on what they will actually do. Hence the next stage in the CPS process: plan formation.

We saw above that a group will not form a collective unless they believe they can actually achieve the desired goal. This, in turn, implies there is at least one action known to the group that will take them 'closer' to the goal (see the definition of J-Able, above). However, it is possible that there are many agents that know of actions the group can perform in order to take them closer to the goal. Moreover, some members of the collective may have objections to one or more of these actions. One of the desiderata for our model, discussed in section 2.1, is that agents are autonomous — they have control over their internal state, and will not simply perform an action because another agent wants them to [33]. It is therefore necessary for the collective to come to some agreement about exactly which course of action they will follow. Such an agreement is reached via *negotiation*.

Negotiation usually involves agents making reasoned arguments for and against courses of action; making proposals and counter proposals; suggesting modifications or amendments to plans; and continuing in this way until all the negotiators have agreed a final result[1]. Negotiation has long been recognised as a process of some importance in multi-agent systems [29, 23]. Unfortunately, these analyses demonstrate that negotiation is also extremely complex — a rigorous attempt at formalization is quite beyond the scope of this paper (see [20] for a logical formalisation of *argumentation*). Instead, we simply offer some observations about the weakest conditions under which negotiation can be said to have occurred.

What can we say about negotiating a plan? First, we note that negotiation may *fail*: the collective may simply be unable to reach agreement, due to some irreconcilable differences. In this case, the minimum condition required for us to be able to say that negotiation occurred at all is that *at least one* agent proposed a course of action which it believed would take the collective closer to the goal. However, negotiation may also succeed. In this case, we expect a team action stage to follow immediately — we shall say no more about team action here, as this is the subject of the next section.

We shall now make the above discussion more precise. First, we define *joint attempts*: what it means for a group of agents to collectively attempt something. As might be expected, joint attempts are a generalisation of single-agent attempts. An attempt by a group of agents g to bring about a state φ is an action α, of which g are the agents, performed with the mutual goal that after α is performed, φ is satisfied, or at least ψ is satisfied (where ψ represents what it takes to make a reasonable effort).

$$\{\textsf{J-Attempt } g\ \alpha\ \varphi\ \psi\} \stackrel{\text{def}}{=} \left[\begin{array}{l} (\textsf{M-Bel } g\ \neg\varphi) \wedge \\ (\textsf{Agts } \alpha\ g) \wedge \\ (\textsf{M-Goal } g\ (\textsf{Achieves } \alpha\ \varphi)) \wedge \\ (\textsf{J-Intend } g\ \alpha; \psi?\ \textsf{true}) \end{array}\right] ?; \alpha$$

We can now state the minimum conditions required for negotiation to have occurred.

[1] It may also involve agents lying, though we shall not consider such cases here.

Intuitively, the group will try to bring about a state where they have agreed on a common plan, and intend to act on it. Failing that, they will bring about a state where at least one of them has proposed a plan which it believed would achieve the desired goal. More formally, if group g are a pre-team with respect to agent i's goal φ, then g will eventually jointly attempt to bring about a state in which g are a team with respect to i's goal φ, or, failing that, to at least bring about a state where some agent $j \in g$, has made g mutually aware of its belief that some action α can be performed by g in order to achieve φ. Formally:

$$(\mathsf{Pre\text{-}Team}\ g\ \varphi\ i) \Rightarrow \mathsf{A}\Diamond\exists\alpha \cdot (\mathsf{Happens}\ \{\mathsf{J\text{-}Attempt}\ g\ \alpha\ p\ q\}) \tag{3}$$

where

$$p \stackrel{\text{def}}{=} (\mathsf{Team}\ g\ \varphi\ i)$$

and

$$q \stackrel{\text{def}}{=} \exists j \cdot \exists\alpha \cdot (j \in g) \wedge (\mathsf{M\text{-}Bel}\ g\ (\mathsf{Bel}\ j\ (\mathsf{Agts}\ \alpha\ g) \wedge (\mathsf{Achieves}\ \alpha\ \varphi))).$$

We can make some other assumptions about agent behaviour during negotiation. Most importantly, we assume that agents will *attempt to bring about their preferences*. For example, if an agent has an objection to some plan, then it will attempt to prevent this plan being carried out. Similarly, if it has a preference for some plan, then it will attempt to bring this plan about. More precisely, if group g are a pre-team with respect to agent i's goal φ, and there is some action α such that it is mutually believed in g that α achieves φ, and that g are the agents of α, then every agent $j \in g$ that has a preference that α does/does not occur will attempt to ensure that α does/does not occur, by at least making g mutually aware of its preference for/against α. Note that we are once again assuming that agents are veracious; that they attempt to influence the team by revealing their true preferences, rather than by lying about their preferences, or not revealing their true preferences.

To formalise the assumption that members make their preferences known, we need to capture the notion of an agent trying to cause and trying to prevent a group performing an action. These are straightforward.

$$\begin{aligned}(\mathsf{Try\text{-}to\text{-}cause}\ i\ g\ \alpha) \stackrel{\text{def}}{=} \exists\alpha' \cdot (\mathsf{Agt}\ \alpha'\ i) \wedge \\ (\mathsf{Happens}\ \{\mathsf{Attempt}\ i\ \alpha'(\mathsf{Does}\ \alpha)\ (\mathsf{M\text{-}Bel}\ g\ (\mathsf{Goal}\ i\ (\mathsf{Does}\ \alpha)))\})\end{aligned}$$

The definition of $(\mathsf{Try\text{-}to\text{-}prevent}\ i\ g\ \alpha)$ is very similar to that of Try-to-cause, and is therefore omitted. The assumption that agents who have a preference for some action make the team mutually aware of their preference is captured in the following assumption.

$$\begin{aligned}\models \forall\alpha \cdot (\mathsf{Pre\text{-}Team}\ g\ \varphi\ i) \wedge (\mathsf{M\text{-}Bel}\ g\ (\mathsf{Agts}\ \alpha\ g) \wedge (\mathsf{Achieves}\ \alpha\ \varphi)) \Rightarrow \\ [\forall j \cdot (j \in g) \wedge (\mathsf{Goal}\ j\ (\mathsf{Does}\ \alpha)) \Rightarrow \mathsf{A}(\mathsf{Try\text{-}to\text{-}cause}\ g\ \alpha)].\end{aligned} \tag{4}$$

Similarly, the assumption that agents who prefer some action not to be performed make the team mutually aware of their preference is captured as follows.

$$\begin{aligned}\models \exists\alpha \cdot (\mathsf{Pre\text{-}Team}\ g\ \varphi\ i) \wedge (\mathsf{M\text{-}Bel}\ g\ (\mathsf{Agts}\ \alpha\ g) \wedge (\mathsf{Achieves}\ \alpha\ \varphi)) \Rightarrow \\ [\forall j \cdot (j \in g) \Rightarrow (\mathsf{Goal}\ j\ (\mathsf{Doesn't}\ \alpha)) \Rightarrow \mathsf{A}(\mathsf{Try\text{-}to\text{-}prevent}\ g\ \alpha)].\end{aligned} \tag{5}$$

If the plan formation phase is successful then the team will have a full joint commitment to the joint goal, and will have agreed to the means by which they will pursue their joint goal[2].

5.4 Team Action

If a collective is successful in its attempt to negotiate a plan, then we expect that collective to follow up negotiation with action. This gives us the fourth, and final stage in our model: team action. For this stage, we simply require that the team has a joint intention of an agreed action. A group g are considered a team with respect to i's goal φ iff there is some action α, such that:

1. α achieves φ; and
2. g have a joint intention of α, relative to i having a goal of φ.

The formalisation of Team is simple.

$$(\mathsf{Team}\ g\ \varphi\ i) \stackrel{\text{def}}{=} \exists \alpha \cdot (\mathsf{M\text{-}Bel}\ g\ (\mathsf{Achieves}\ \alpha\ \varphi)) \wedge (\mathsf{J\text{-}Intend}\ g\ \alpha\ (\mathsf{Goal}\ i\ \varphi)).$$

At this stage, the commitment that agent have to action is essentially that characterised by Levesque-Cohen-Nunes in their model of teamwork [18]. From the definition of J-Intend, we know that the group will remain committed to mutually believing they are about to perform the action, and then performing it. Moreover, if ever one of them comes to believe, for example, that i no longer has a goal of φ, then this agent will make the team aware of this, and team action will end. It is straightforward to prove the following properties of team action.

Theorem 21 $\models (\mathsf{Team}\ g\ \varphi\ i) \Rightarrow (\mathsf{M\text{-}Bel}\ g\ \mathsf{E}\Diamond\varphi)$

Theorem 22 $\models (\mathsf{Team}\ g\ \varphi\ i) \Rightarrow (\mathsf{M\text{-}Bel}\ g\ \mathsf{E}(\mathsf{Happens}\ \alpha))$

So a team working toward an agent i's goal φ mutually believe that φ is possible. Moreover, they mutually believe that the action they intend to perform in order to achieve φ can actually happen.

6 Desiderata Revisited

In section 2.1 we identified a number of properties that an adequate theory of CPS should exhibit. We now revisit these properties, and see how our model stands up against them.

- *Agents are autonomous.*

[2] Ideally, we would like to specify that the group also negotiate a *convention* for monitoring team action. Unfortunately, we have no direct way of representing such behaviour: it would require quantification over formulae of the language, and such a meta-level notion cannot be represented at the object level in a normal modal language such as that used here.

The model predicts that once the agents are formed into a collective, they will attempt to negotiate a plan that they believe will achieve the desired objective. Moreover, they will make their preferences known with respect to such plans, and are not required simply to accept another agent's proposal; they are therefore autonomous.

- *Cooperation can fail.*

  There are a number of stages at which the cooperation process may fail. First, an agent that has recognised the potential for cooperation may be unable to form a team of agents. Secondly, having formed the team, the agents may be unable to agree upon a plan of action. Finally, cooperation may fail after a plan has been agreed because of unforeseen circumstances or because one of the agents drops its commitment to the endeavour.

- *Communication is essential.*

  Although we have not explicitly considered communication, our model is consistent with one of the best current theories of speech acts: in [6], Cohen-Levesque built a theory in which illocutionary acts are treated as *attempts* to bring about some mental state in conversation participants [6, p227,pp240–241]. At a number of points, our model predicts precisely such attempts. For example, in the team formation stage, an agent that recognises the potential for cooperation will perform some action in an attempt to bring about a Pre-Team mental state in some group that it believes can help with its goal.

- *Communication acts are characterised by their effects.*

  In our model, rational agents will communicate with other agents if they recognise the potential for cooperation with respect to one of their goals. However, our model does not require that agents use any pre-defined communication language or cooperation protocol. In our model, as in that of [6], *any* action can be viewed as communicative, as long as it is performed by an agent in the appropriate circumstances.

- *Agents are reactive.*

  The model presented above is essentially a set of *liveness* properties [21]; this is consistent with the view of agents as *intelligent reactive systems*, responding in a reasoned way to their goals, and events that occur in their environment. Moreover, the agents have a specific set of conditions and associated goals specified in their convention, which indicate the events they should respond to.

- *Agents initiate social processes.*

  The model predicts that agents will attempt to initiate social interaction if they have some goals which they cannot achieve in isolation or for which they prefer the assistance of others. Moreover, agents will initiate the social process of team planning if they reach the state of being a Pre-Team, and the social process of team planning if they reach the state of being a team.

- *Agents will be mutually supportive.*

During the planning phase, the agents support one-another by making sure that they inform their fellow group members if they believe the plan will not achieve its intended aim (for whatever reason). During execution, the social convention ensures that agents support one-another by ensuring that others know when they believe the cooperative activity is in difficulty.

7 Discussion

This article has contributed to the theoretical foundations of multi-agent systems by presenting a formal model of the cooperative problem solving process. This four-stage model predicts and describes the circumstances under which agents will recognise the potential for cooperation, and how they will behave when this situation arises, from attempting to build a team, negotiating a collective plan, and acting as a team. We noted that this model is both abstract and idealised: there are cases that it does not consider, and no doubt some assumptions have been made that are either too strong or too weak. Nevertheless, we are aware of no other attempt to formalise the cooperative problem solving process in this way.

The fundamental nature and form of the model was deliberately chosen to provide assistance to practitioners who are concerned with developing cooperating agents. The model provides a coherent set of conceptual mechanisms upon which cooperative behaviour can be based. Thus, these mechanisms can be used to identify a cooperating agent's key data structures, the properties that these structures should exhibit, the operations which can be performed on the structures, and the various inter-relationships which exist between the structures. Such models are especially useful when the cooperating agent is to be realised using a (traditional) symbolic AI architecture since there is a comparatively straightforward mapping between the model and the architecture's separation of concerns.

There are a number of issues that we intend to address in future work, the most obvious of which is the need for refinement of the model, including more detailed treatments of the process of recognising potential for cooperation; the process of building a team; the process of negotiation; and the various conventions that may be used for collective action. Finally, we have said nothing about the meta-cooperative process by which agents come to agree on a convention itself: we have taken conventions as given. In real-world cooperative scenarios, such activities are just one part of the cooperative process, that must be addressed by researchers in multi-agent systems.

Acknowledgements

The second author was supported by a grant from the Queen Mary & Westfield College Engineering Faculty Research Fund. Prototypical versions of this paper appeared as [31, 32].

A The Formal Framework: A Complete Definition

A.1 Syntax

Definition 23 The language contains the following symbols:

1. the *propositional connectives* $\neg$ (not) and $\vee$ (or), and *universal quantifier* $\forall$;
2. the *operator symbols* Bel, Goal, Happens, Agts, $\in$, $=$, and A;
3. a countable set $Pred$ of *predicate symbols* — each symbol $P \in Pred$ is associated with a natural number called its *arity*, given by $arity(P)$;
4. a countable set $Const$ of *constant symbols*, the union of the mutually disjoint sets $Const_{Ag}$ (agent constants), $Const_{Ac}$ (action sequence constants), $Const_{Gr}$ (group constants), and $Const_U$ (other constants);
5. a countable set Var of *variable symbols*, the union of the mutually disjoint sets Var_{Ag}, Var_{Ac}, Var_{Gr} and Var_U;
6. the action expression constructors ';', '|', '$*$', and '?';
7. the *punctuation symbols*), (, $\cdot$ and comma ','.

Definition 24 A *term* is either a constant or a variable; the set of terms is $Term$. The *sort* of a term is either Ag, Ac, Gr or U; if s is a sort then by $Term_s$ we mean $Const_s \cup Var_s$. Thus $\tau_s \in Term_s$.

Notice that the language contains constants, but no other functional terms. The syntax of (well-formed) formulae ($\langle fmla \rangle$) of the language is defined in Figure 1. Note that we demand that a predicate P is applied to $arity(P)$ terms.

A.2 Semantics

First, some general concepts. It is assumed that the world may be in any of a set S of *states*. A state *transition* is caused by the occurrence of a *primitive action* (or *event*): the set of all primitive actions is D_{Ac}. From any state, there is at least one — and perhaps many — possible actions, and hence resultant states. The binary relation R on S is used to represent all possible courses of world history: $(s, s') \in R$ iff the state s could be transformed into state s' by the occurrence of a primitive action that is possible in s. Clearly, R will *branch* infinitely into the future from every state. A labelling function Act maps each arc in R to the action associated with the transition.

The world is populated by a non-empty set D_{Ag} of *agents*. A *group* over D_{Ag} is simply a non-empty subset of D_{Ag}; the set of all such groups is D_{Gr}. Agents and groups may easily be related to one-another via a simple (typed) set theory. Agents have beliefs and goals, and are (idealised) reasoners. The beliefs of an agent are given by a *belief accessibility relation* on S in the usual way; similarly for goals. Every primitive action α is associated with an agent, given by $Agt(\alpha)$. Finally, the world contains other individuals (chairs, pints of beer, etc.) given by the set D_U. A complete formal definition of the language semantics will now be given. First, *paths* (a.k.a. fullpaths) will be defined: a path represents a possible course of events through a branching time structure.

Definition 25 If S is a non-empty set and R is a total binary relation on S then a *path* over S, R is an infinite sequence $(s_u : u \in \mathbb{N})$ such that $\forall u \in \mathbb{N}$, $s_u \in S$ and $(s_u, s_{u+1}) \in R$. The set of all paths over S, R is given by $paths(S, R)$. The *head* of a path $p = (s_0, \ldots)$ is its first element s_0, and is given by $hd(p)$.

$$
\begin{array}{rcl}
\langle \textit{ag-term} \rangle & ::= & \text{any element of } Term_{Ag} \\
\langle \textit{ac-term} \rangle & ::= & \text{any element of } Term_{Ac} \\
\langle \textit{gr-term} \rangle & ::= & \text{any element of } Term_{Gr} \\
\langle \textit{term} \rangle & ::= & \text{any element of } Term \\
\langle \textit{pred-sym} \rangle & ::= & \text{any element of } Pred \\
\langle \textit{var} \rangle & ::= & \text{any element of } Var \\
\langle \textit{ac-exp} \rangle & ::= & \langle \textit{ac-term} \rangle \\
 & | & \langle \textit{ac-exp} \rangle \ ; \ \langle \textit{ac-exp} \rangle \\
 & | & \langle \textit{ac-exp} \rangle \ \text{`|'} \ \langle \textit{ac-exp} \rangle \\
 & | & \langle \textit{state-fmla} \rangle ? \\
 & | & \langle \textit{ac-exp} \rangle * \\
\langle \textit{state-fmla} \rangle & ::= & \langle \textit{pred-sym} \rangle (\langle \textit{term} \rangle, \cdots, \langle \textit{term} \rangle) \\
 & | & (\mathsf{Bel} \ \langle \textit{ag-term} \rangle \ \langle \textit{state-fmla} \rangle) \\
 & | & (\mathsf{Goal} \ \langle \textit{ag-term} \rangle \ \langle \textit{state-fmla} \rangle) \\
 & | & (\mathsf{Agts} \ \langle \textit{ac-term} \rangle \ \langle \textit{gr-term} \rangle) \\
 & | & (\langle \textit{term} \rangle = \langle \textit{term} \rangle) \\
 & | & (\langle \textit{ag-term} \rangle \ \in \ \langle \textit{gr-term} \rangle) \\
 & | & \mathsf{A} \langle \textit{path-fmla} \rangle \\
 & | & \neg \langle \textit{state-fmla} \rangle \\
 & | & \langle \textit{state-fmla} \rangle \vee \langle \textit{state-fmla} \rangle \\
 & | & \forall \langle \textit{var} \rangle \cdot \langle \textit{state-fmla} \rangle \\
\langle \textit{path-fmla} \rangle & ::= & (\mathsf{Happens} \ \langle \textit{ac-exp} \rangle) \\
 & | & \langle \textit{state-fmla} \rangle \\
 & | & \neg \langle \textit{path-fmla} \rangle \\
 & | & \langle \textit{path-fmla} \rangle \vee \langle \textit{path-fmla} \rangle \\
 & | & \forall \langle \textit{var} \rangle \cdot \langle \textit{path-fmla} \rangle \\
\langle \textit{fmla} \rangle & ::= & \langle \textit{state-fmla} \rangle
\end{array}
$$

Figure 1: Syntax

Next, we present the technical apparatus for dealing with the denotation of terms.

Definition 26 The *domain of quantification*, D, is $D_{Ag} \cup (D_{Ac}^*) \cup D_{Gr} \cup D_U$, (where S^* denotes the set of non-empty sequences over S). If $n \in I\!N$, then the set of n-tuples over D is denoted by D^n.

The language thus allows quantification over agents, sequences of primitive actions, groups, and other individuals. (Note that D is fixed, for all states.)

Definition 27 An *interpretation for constants*, I, is a sort-preserving bijection $I : Const \to D$. A *variable assignment*, V, is a sort-preserving bijection $V : Var \to D$.

Constants are therefore rigid designators. It is possible to derive a function which returns the denotation of an arbitrary term relative to I, V.

Definition 28

$$[\![\tau]\!]_{I,V} \stackrel{\text{def}}{=} \begin{cases} I(\tau) & \text{if } \tau \in Const \\ V(\tau) & \text{otherwise.} \end{cases}$$

Reference to I, V will usually be suppressed.

Definition 29 A *model*, M, is a structure:

$$\langle S, R, D_{Ag}, D_{Ac}, D_{Gr}, D_U, Act, Agt, B, G, I, \Phi \rangle$$

where:

- S is a non-empty set of states;
- $R \subseteq S \times S$ is a total binary relation on S;
- D_{Ag} is a non-empty set of agents;
- D_{Ac} is a non-empty set of actions;
- D_{Gr} is the set of groups over D_{Ag};
- D_U is a non-empty set of other individuals;
- $Act : R \to D_{Ac}$ associates a primitive action with each arc in R;
- $Agt : D_{Ac} \to D_{Ag}$ gives the agent of each primitive action;
- $B : D_{Ag} \to \wp(S \times S)$ associates a transitive, Euclidean, serial belief accessibility relation with every agent in D_{Ag};
- $G : D_{Ag} \to \wp(S \times S)$ associates a serial goal accessibility relation with every agent in D_{Ag}, such that:
  1. $\forall i \in D_{Ag}$, $G(i) \subseteq B(i)$;
  2. $\forall i \in D_{Ag}$, if $(s, s') \in G(i)$ and $(s, s'') \in B(i)$ then $(s'', s') \in G(i)$;
- $I : Const \to D$ is an interpretation for constants; and finally
- Φ is a function

  $$\Phi : Pred \times S \to \bigcup_{n \in \mathbb{N}} D^n$$

  which gives the extension of each predicate symbol in each state, such that

  $$\forall P \in Pred, \forall n \in \mathbb{N}, \forall s \in S, \text{ if } arity(P) = n \text{ then } \Phi(P, s) \subseteq D^n.$$

  (i.e., it preserves arity).

$$
\begin{array}{rcl}
occurs(\alpha,u,v,(s_0,\ldots)) & \text{iff} & [\![\alpha]\!] = (\alpha_1,\ldots,\alpha_n), n \leq v-u \text{ and } \forall w \in \{1,\ldots,n\}, \\
 & & \quad Act(s_{u+w-1}, s_{u+w}) = \alpha_w \\
 & & \quad (\text{where } \alpha \in Term_{Ac}) \\
occurs(\alpha;\alpha',u,v,p) & \text{iff} & \exists w \in \{u,\ldots,v\} \text{ s.t. } occurs(\alpha,u,w,p) \text{ and} \\
 & & \quad occurs(\alpha',w,v,p) \\
occurs(\alpha|\alpha',u,v,p) & \text{iff} & occurs(\alpha,u,v,p) \text{ or } occurs(\alpha',u,v,p) \\
occurs(\varphi?,u,v,p) & \text{iff} & \langle M,V,hd(p)\rangle \models \varphi \\
occurs(\alpha*,u,v,p) & \text{iff} & \exists w_1,\ldots,w_x \in I\!N \text{ s.t. } (w_1 = 0) \text{ and } (w_1 < \cdots < w_x) \\
 & & \quad \text{and } \forall y \in \{1,\ldots,x\}, occurs(\alpha,w_y,w_{y+1},p)
\end{array}
$$

Figure 2: The meta-language '*occurs*' predicate

The semantics of the language are defined via the satisfaction relation, '$\models$', which holds between *interpretation structures* and formulae of the language. For state formulae, an interpretation structure is a triple $\langle M,V,s\rangle$, where M is a model, V is a variable assignment and s is a state. For path formulae, an interpretation structure is a triple $\langle M,V,p\rangle$, where p is a path. The rules defining the satisfaction relation are given in Figure 3 (state formulae) and Figure 4 (path formulae). The rules make use of some syntactic abbreviations. First, we write $occurs(\alpha,u,v,p)$ if action α occurs between 'times' $u,v \in I\!N$ on the (possibly finite) path p: this meta-level predicate is defined by the rules in Figure 2.

Additionally, two functions are defined that return all the primitive actions referred to in an action sequence, and the agents required for an action term, respectively.

$$
\begin{array}{rcl}
actions((\alpha_1,\ldots,\alpha_n)) & \stackrel{\text{def}}{=} & \{\alpha_1,\ldots,\alpha_n\} \\
agents(\alpha) & \stackrel{\text{def}}{=} & \{i \mid \exists \alpha' \in actions([\![\alpha]\!]) \text{ s.t. } Agt(\alpha') = i\} \\
 & & \quad (\text{where } \alpha \in Term_{Ac})
\end{array}
$$

A.3 Some Properties

If a formula φ is *valid* (satisfied by all interpretation structures), we write $\models \varphi$, as usual. The language defined above is a many-sorted first-order language, and it inherits the expected properties of such languages. Additionally, the Bel and Goal operators have the properties that one would expect of them, given the restrictions on accessibility relations enforced above: the logic of Bel is KD45, and the logic of Goal is KD. Necessitation works for both Bel and Goal [4].

Theorem 30

$$
\models \forall i \cdot (\mathsf{Bel}\ i\ \varphi \Rightarrow \psi) \quad \Rightarrow \quad ((\mathsf{Bel}\ i\ \varphi) \Rightarrow (\mathsf{Bel}\ i\ \psi)) \tag{6}
$$

$$
\models \forall i \cdot (\mathsf{Bel}\ i\ \varphi) \quad \Rightarrow \quad \neg(\mathsf{Bel}\ i\ \neg\varphi) \tag{7}
$$

$$
\models \forall i \cdot (\mathsf{Bel}\ i\ \varphi) \quad \Rightarrow \quad (\mathsf{Bel}\ i\ (\mathsf{Bel}\ i\ \varphi)) \tag{8}
$$

$$
\begin{array}{lcll}
\langle M,V,s\rangle & \models & \mathsf{true} & \\
\langle M,V,s\rangle & \models & P(\tau_1,\ldots,\tau_n) & \text{iff } \langle [\![\tau_1]\!],\ldots,[\![\tau_n]\!]\rangle \in \Phi(P,s) \\
\langle M,V,s\rangle & \models & (\mathsf{Bel}\ i\ \varphi) & \text{iff } \forall s' \in S,\ \text{if } (s,s') \in B([\![i]\!]) \\
 & & & \qquad \text{then } \langle M,V,s'\rangle \models \varphi \\
\langle M,V,s\rangle & \models & (\mathsf{Goal}\ i\ \varphi) & \text{iff } \forall s' \in S,\ \text{if } (s,s') \in G([\![i]\!]) \\
 & & & \qquad \text{then } \langle M,V,s'\rangle \models \varphi \\
\langle M,V,s\rangle & \models & (\mathsf{Agts}\ \alpha\ g) & \text{iff } agents(\alpha) = [\![g]\!] \\
\langle M,V,s\rangle & \models & (\tau_1 = \tau_2) & \text{iff } [\![\tau_1]\!] = [\![\tau_2]\!] \\
\langle M,V,s\rangle & \models & (i \in g) & \text{iff } [\![i]\!] \in [\![g]\!] \\
\langle M,V,s\rangle & \models & \mathsf{A}\varphi & \text{iff } \forall p \in paths(S,R),\ \text{if } hd(p) = s \\
 & & & \qquad \text{then } \langle M,V,p\rangle \models \varphi \\
\langle M,V,s\rangle & \models & \neg\varphi & \text{iff } \langle M,V,s\rangle \not\models \varphi \\
\langle M,V,s\rangle & \models & \varphi \vee \psi & \text{iff } \langle M,V,s\rangle \models \varphi \text{ or } \langle M,V,s\rangle \models \psi \\
\langle M,V,s\rangle & \models & \forall x \cdot \varphi & \text{iff } \langle M, V \dagger \{x \mapsto d\}, s\rangle \models \varphi \\
 & & & \qquad \text{for all } d \in D \text{ s.t. } x \text{ and } d \text{ are} \\
 & & & \qquad \text{of the same sort}
\end{array}
$$

Figure 3: State Formulae Semantics

$$
\begin{array}{rcll}
\models \forall i \cdot \neg(\mathsf{Bel}\ i\ \varphi) & \Rightarrow & (\mathsf{Bel}\ i\ \neg(\mathsf{Bel}\ i\ \varphi)) & (9) \\
\models \varphi & \rightarrow & \models \forall i \cdot (\mathsf{Bel}\ i\ \varphi) & (10) \\
\models \forall i \cdot (\mathsf{Goal}\ i\ \varphi \Rightarrow \psi) & \Rightarrow & ((\mathsf{Goal}\ i\ \varphi) \Rightarrow (\mathsf{Goal}\ i\ \psi)) & (11) \\
\models \forall i \cdot (\mathsf{Goal}\ i\ \varphi) & \Rightarrow & \neg(\mathsf{Goal}\ \neg\varphi) & (12) \\
\models \varphi & \rightarrow & \models \forall i \cdot (\mathsf{Goal}\ i\ \varphi) & (13)
\end{array}
$$

Proof These properties are generalisations of the corresponding modal logic theorems implied by the restrictions imposed on the model theory of the language — see, e.g., [4].

Turning to the relationship between beliefs and goals, we can prove the following.

Theorem 31

$$
\begin{array}{rcll}
\models \forall i \cdot (\mathsf{Bel}\ i\ \varphi) & \Rightarrow & (\mathsf{Goal}\ i\ \varphi) & (14) \\
\models \forall i \cdot (\mathsf{Goal}\ i\ \varphi) & \Rightarrow & \neg(\mathsf{Bel}\ i\ \neg\varphi) & (15) \\
\models \forall i \cdot (\mathsf{Goal}\ i\ \varphi) & \Rightarrow & (\mathsf{Bel}\ i\ (\mathsf{Goal}\ i\ \varphi)) & (16)
\end{array}
$$

Proof Axiom (14) is known as *realism*, and is a consequence of an agent's goal accessibility relation being a subset of its belief accessibility relation [5, pp227–228]. The second realism axiom (15) also follows from this constraint. For suppose that $\langle M,V,s\rangle \models (\mathsf{Goal}\ i\ \varphi)$, for arbitrary $\langle M,V,s\rangle$. Then for all $s' \in S$ such that $(s,s') \in G([\![i]\!])$, we have $\langle M,V,s'\rangle \models \varphi$, and since $G([\![i]\!]) \subseteq B([\![i]\!])$, it must be that $(s,s') \in B([\![i]\!])$. Hence $\langle M,V,s\rangle \models \neg(\mathsf{Bel}\ i\ \neg\varphi)$. For (16), assume $\langle M,V,s\rangle \models (\mathsf{Goal}\ i\ \varphi)$

$\langle M, V, p\rangle$	$\models$	$(\mathsf{Happens}\ \alpha)$	iff $\exists u \in I\!N$ s.t. $occurs(\alpha, 0, u, p)$
$\langle M, V, p\rangle$	$\models$	φ	iff $\langle M, V, hd(p)\rangle \models \varphi$ (where φ is a state formula)
$\langle M, V, p\rangle$	$\models$	$\neg\varphi$	iff $\langle M, V, p\rangle \not\models \varphi$
$\langle M, V, p\rangle$	$\models$	$\varphi \vee \psi$	iff $\langle M, V, p\rangle \models \varphi$ or $\langle M, V, p\rangle \models \psi$
$\langle M, V, p\rangle$	$\models$	$\forall x \cdot \varphi$	iff $\langle M, V \dagger \{x \mapsto d\}, p\rangle \models \varphi$ for all $d \in D$ s.t. x and d are of the same sort

Figure 4: Path Formulae Semantics

for arbitrary φ. Hence $\langle M, V, s'\rangle \models \varphi$ for all $s' in S$ such that $(s, s') \in G([\![i]\!])$. We need to show that for all $s'' \in S$ such that $(s, s'') \in B([\![i]\!])$, we have $\langle M, V, s''\rangle \models (\mathsf{Goal}\ i\ \varphi)$. But if $(s, s') \in G([\![i]\!])$ and $(s, s'') \in B([\![i]\!])$, then from the constraint on the goal relation, we must have $(s'', s') \in G([\![i]\!])$, and since $\langle M, V, s'\rangle \models \varphi$, we have $\langle M, V, s''\rangle \models (\mathsf{Goal}\ i\ \varphi)$, and we are done.

We will also require that the logic satisfies the following *strong realism* constraint.

$$\models \forall i \cdot (\mathsf{Goal}\ i\ \mathsf{E}\varphi) \Rightarrow (\mathsf{Bel}\ i\ \mathsf{E}\varphi) \tag{17}$$

Thus if an agent has a goal that φ is possibly satisfied, then it believes that φ is possibly satisfied. The semantic constraint corresponding to this axiom is quite intuitive, but we omit it in the interests of brevity — the reader is referred to [22, pp317–333] for a discussion.

Theorem 32

$$\models \mathsf{A}(\varphi \Rightarrow \psi) \quad \Rightarrow \quad ((\mathsf{A}\varphi) \Rightarrow (\mathsf{A}\psi)) \tag{18}$$

$$\models \mathsf{A}\varphi \quad \Rightarrow \quad \varphi \tag{19}$$

$$\models \neg\mathsf{A}\varphi \quad \Rightarrow \quad \mathsf{A}\neg\mathsf{A}\varphi \tag{20}$$

$$\models \varphi \quad \rightarrow \quad \models \mathsf{A}\varphi \tag{21}$$

The A operator thus has the properties of a normal modal operator based on a universal relation, and thus analogues of the modal axioms KT5 (modal system S5) hold for this operator [4, p98]; also, a version on necessitation holds.

The following theorem captures some simple properties of action expressions and the $\mathsf{Happens}$ operator that are used in our proofs (see [5, p229] for others).

Theorem 33

$$\models (\mathsf{Happens}\ \varphi?) \quad \Rightarrow \quad \varphi \tag{22}$$

$$\models (\mathsf{Happens}\ \alpha; \varphi?) \quad \Rightarrow \quad \Diamond\varphi \tag{23}$$

Proof For (22), assume $\langle M, V, p\rangle \models (\mathsf{Happens}\ \varphi?)$ for arbitrary p. Then by the semantics of Happens, we have $\langle M, V, hd(p)\rangle \models \varphi$, hence $\langle M, V, P\rangle \models \varphi$. For (23), assume $\langle M, V, p\rangle \models (\mathsf{Happens}\ \alpha; \varphi?)$. Then by the semantics of Happens, we have that $occurs(\varphi?, u, v, p)$ for some $u, v \in I\!N$, such that $v > u$. Hence $\langle M, V, p\rangle \models \Diamond\varphi$.

The M-Bel operator has properties rather similar to those of Bel.

Theorem 34

$$\models \forall g \cdot (\mathsf{M\text{-}Bel}\ g\ \varphi \Rightarrow \psi) \quad \Rightarrow \quad ((\mathsf{M\text{-}Bel}\ g\ \varphi) \Rightarrow (\mathsf{M\text{-}Bel}\ g\ \psi)) \tag{24}$$

$$\models \forall g \cdot (\mathsf{M\text{-}Bel}\ g\ \varphi) \quad \Rightarrow \quad \neg(\mathsf{M\text{-}Bel}\ g\ \neg\varphi) \tag{25}$$

$$\models \forall g \cdot (\mathsf{M\text{-}Bel}\ g\ \varphi) \quad \Rightarrow \quad (\mathsf{M\text{-}Bel}\ g\ (\mathsf{M\text{-}Bel}\ g\ \varphi)) \tag{26}$$

$$\models \forall g \cdot \neg(\mathsf{M\text{-}Bel}\ g\ \varphi) \quad \Rightarrow \quad (\mathsf{M\text{-}Bel}\ g\ \neg(\mathsf{M\text{-}Bel}\ g\ \varphi)) \tag{27}$$

$$\models \varphi \quad \rightarrow \quad \models (\mathsf{M\text{-}Bel}\ g\ \varphi) \tag{28}$$

References

[1] R. Axelrod. *The Evolution of Cooperation.* Basic Books, 1984.

[2] K. Binmore. *Fun and Games: A Text on Game Theory.* D. C. Heath and Company: Lexington, MA, 1992.

[3] M. E. Bratman. Planning and the stability of intentions. *Minds and Machines*, 2:1–16, 1992.

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

[5] P. R. Cohen and H. J. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42:213–261, 1990.

[6] P. R. Cohen and H. J. Levesque. Rational interaction as the basis for communication. In P. R. Cohen, J. Morgan, and M. E. Pollack, editors, *Intentions in Communication*, pages 221–256. The MIT Press: Cambridge, MA, 1990.

[7] E. H. Durfee. *Coordination of Distributed Problem Solvers.* Kluwer Academic Publishers: Boston, MA, 1988.

[8] E. A. Emerson and J. Y. Halpern. 'Sometimes' and 'not never' revisited: on branching time versus linear time temporal logic. *Journal of the ACM*, 33(1):151–178, 1986.

[9] R. Fagin, J. Y. Halpern, Y. Moses, and M. Y. Vardi. *Reasoning About Knowledge.* The MIT Press: Cambridge, MA, 1995.

[10] J. R. Galliers. *A Theoretical Framework for Computer Models of Cooperative Dialogue, Acknowledging Multi-Agent Conflict.* PhD thesis, Open University, UK, 1988.

[11] A. Haddadi. *Communication and Cooperation in Agent Systems (LNAI Volume 1056)*. Springer-Verlag: Berlin, Germany, 1996.

[12] J. Y. Halpern and Y. Moses. A guide to completeness and complexity for modal logics of knowledge and belief. *Artificial Intelligence*, 54:319–379, 1992.

[13] J. Y. Halpern and M. Y. Vardi. The complexity of reasoning about knowledge and time. I. Lower bounds. *Journal of Computer and System Sciences*, 38:195–237, 1989.

[14] D. Harel. Dynamic logic. In D. Gabbay and F. Guenther, editors, *Handbook of Philosophical Logic Volume II — Extensions of Classical Logic*, pages 497–604. D. Reidel Publishing Company: Dordrecht, The Netherlands, 1984. (Synthese library Volume 164).

[15] G. E. Hughes and M. J. Cresswell. *Introduction to Modal Logic*. Methuen and Co., Ltd., 1968.

[16] N. R. Jennings. Commitments and conventions: The foundation of coordination in multi-agent systems. *The Knowledge Engineering Review*, 8(3):223–250, 1993.

[17] N. R. Jennings. Controlling cooperative problem solving in industrial multi-agent systems using joint intentions. *Artificial Intelligence*, 75:195–240, 1995.

[18] H. J. Levesque, P. R. Cohen, and J. H. T. Nunes. On acting together. In *Proceedings of the Eighth National Conference on Artificial Intelligence (AAAI-90)*, pages 94–99, Boston, MA, 1990.

[19] R. C. Moore. A formal theory of knowledge and action. In J. F. Allen, J. Hendler, and A. Tate, editors, *Readings in Planning*, pages 480–519. Morgan Kaufmann Publishers: San Mateo, CA, 1990.

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

[21] A. Pnueli. Specification and development of reactive systems. In *Information Processing 86*. Elsevier Science Publishers B.V.: Amsterdam, The Netherlands, 1986.

[22] A. S. Rao and M. Georgeff. Decision procedures of BDI logics. *Journal of Logic and Computation*, 8(3):293–344, 1998.

[23] J. S. Rosenschein and G. Zlotkin. *Rules of Encounter: Designing Conventions for Automated Negotiation among Computers*. The MIT Press: Cambridge, MA, 1994.

[24] S. J. Russell. Rationality and intelligence. *Artificial Intelligence*, 94(1-2):57–78, July 1997.

[25] J. R. Searle. Collective intentions and actions. In P. R. Cohen, J. Morgan, and M. E. Pollack, editors, *Intentions in Communication*, pages 401–416. The MIT Press: Cambridge, MA, 1990.

[26] M. P. Singh. The intentions of teams: Team structure, endodeixis, and exodeixis. In *Proceedings of the Thirteenth European Conference on Artificial Intelligence (ECAI-98)*, pages 303–307, Brighton, United Kingdom, 1998.

[27] R. G. Smith. *A Framework for Distributed Problem Solving.* UMI Research Press, 1980.

[28] L. Steels. Cooperation between distributed agents through self organization. In Y. Demazeau and J.-P. Müller, editors, *Decentralized AI — Proceedings of the First European Workshop on Modelling Autonomous Agents in a Multi-Agent World (MAAMAW-89)*, pages 175–196. Elsevier Science Publishers B.V.: Amsterdam, The Netherlands, 1990.

[29] K. P. Sycara. Multiagent compromise via negotiation. In L. Gasser and M. Huhns, editors, *Distributed Artificial Intelligence Volume II*, pages 119–138. Pitman Publishing: London and Morgan Kaufmann: San Mateo, CA, 1989.

[30] M. Wooldridge and M. Fisher. A first-order branching time logic of multi-agent systems. In *Proceedings of the Tenth European Conference on Artificial Intelligence (ECAI-92)*, pages 234–238, Vienna, Austria, 1992.

[31] M. Wooldridge and N. R. Jennings. Formalizing the cooperative problem solving process. In *Proceedings of the Thirteenth International Workshop on Distributed Artificial Intelligence (IWDAI-94)*, pages 403–417, Lake Quinalt, WA, July 1994.

[32] M. Wooldridge and N. R. Jennings. Towards a theory of cooperative problem solving. In *Proceedings of the Sixth European Workshop on Modelling Autonomous Agents and Multi-Agent Worlds (MAAMAW-94)*, pages 15–26, August 1994.

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