Towards a Theory of Cooperative Problem Solving

Michael Wooldridge

Dept. of Computing
Manchester Metropolitan University
Chester Street, Manchester M1 5GD
United Kingdom

M.Wooldridge@doc.mmu.ac.uk

Nicholas R. Jennings

Dept. of Electronic Engineering
Queen Mary & Westfield College
Mile End Road, London E1 4NS
United Kingdom

N.R.Jennings@qmw.ac.uk

Abstract. One objective of distributed artificial intelligence research is to build systems that are capable of cooperative problem solving. To this end, a number of implementation-oriented models of cooperative problem solving have been developed. However, *mathematical* models of social activity have focussed only on limited aspects of the cooperative problem solving process: no mathematical model of the entire process has yet been described. In this paper, we rectify this omission. We present a preliminary model that describes the cooperative problem solving process from recognition of the potential for cooperation through to team action. The model is formalised by representing it as a theory in a quantified multi-modal logic. A key feature of the model is its reliance on the twin notions of *commitments* and *conventions*; conventions (protocols for monitoring commitments) are formalised for the first time in this paper. We comment on the generality of the model, outline its deficiencies, and suggest some possible refinements and other future areas of research.

1 Introduction

Distributed Artificial Intelligence (DAI) is concerned with all forms of social activity in systems composed of multiple computational agents [1]. An important form of interaction in such systems is *cooperative problem solving* (CPS), which occurs when a group of logically decentralised agents choose to work together to achieve a common goal. Relevant examples include a group of agents moving a heavy object, playing a symphony, building a house, and writing a joint paper. As these examples indicate, CPS is a common and important process in human societies, and there is increasing evidence to support the claim that it will be similarly important in future computer systems. A number of models of the CPS process have been devised by DAI researchers. Some of these models represent frameworks for implementing CPS systems, and for managing cooperative activities in such systems at run-time (e.g., [15, 5]). Other, more formal models have been developed in an attempt to characterise various aspects of CPS (e.g., [10, 8, 17]).

As is the case in mainstream AI, the differing motivations and approaches of formalists and system builders has meant that there has been little cross-fertilisation between the two areas. The former camp has concentrated on isolated aspects of the CPS process,

whereas work in the latter camp has concentrated on devising protocols for the entire CPS process. However, the key assumptions and design decisions of implementation-oriented CPS models tend to be buried deep inside the associated software; this can make it difficult to extract general principles or results from implementations.

This paper goes some way to bridging the gap between theory and practice in DAI. We develop a four-stage model of CPS, which we make precise by expressing it as a theory in a quantified multi-modal logic. The development of this model was driven by an analysis of CPS in both natural and artificial systems; the result is a theory that is accessible to both formalists and system builders. For formalists, the model represents a first attempt to capture the properties of CPS in a mathematical framework, with the corollary that properties of the model may be established via formal proof. For system builders, the model can serve as an abstract, top-level specification of a CPS system, which can inform the development of future DAI applications. The model deals with a number of issues that have hitherto been neglected by DAI theorists; for example, it considers the process by which an agent recognises the potential for cooperation, and begins to solicit assistance. Note that although we have attempted to develop a model that deals with CPS from beginning to end, we do not claim that our model is the final word on the subject; it would not be possible to present, in such a short paper, a theory that dealt with all conceivable aspects of a process as complex as CPS (see §5).

The remainder of this paper is structured as follows. The following section presents an overview of the formal framework used to represent the model. In §3, the notions of commitments and conventions, which play a key role in our model, are discussed and subsequently formalised; the model of CPS is then developed in §4. Some conclusions are presented in §5.

2 A Formal Framework

This section gives an overview of the formal framework in which the model of CPS will be expressed. This framework is a quantified multi-modal logic, which both draws upon and extends the work described in [3, 13]. Unfortunately, space restrictions prevent us from defining the language in full here; a complete formal definition of the language's syntax and semantics may be found in [17].

Informally, the operators of the language have the following meanings. The operator true is a logical constant for truth. (Bel i φ) and (Goal i φ) mean that agent i has a belief, or goal of φ respectively. The = operator is usual first-order equality. 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. The (Agts α g) operator means that the group denoted by g are precisely the agents required to perform the actions in the action sequence denoted by α. The A operator is a *path quantifier*: Aφ means that φ is a *path formula* that is satisfied in all the futures that could arise from the current state[1]. The operators $\neg$ (not) and $\vee$ (or)

[1] There is a distinction made in the language between *path* and *state* formulae: state formulae are evaluated with respect to the 'current state' of the world, whereas path formulae are evaluated with respect to a course of events. The well-formed formulae of the language are identified with the set of state formulae [6].

have classical semantics, as does the universal quantifier $\forall$; the remaining classical connectives and existential quantifier are assumed to be introduced as abbreviations, in the obvious way. (Happens α) is a path formula that means that the action α happens next; $\alpha; \alpha'$ means the action α is immediately followed by α'; $\alpha|\alpha'$ means either α or α' happen next; $\varphi?$ is a test action, which occurs if φ is 'true' in the current state; $\alpha*$ means the action α iterated.

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 $\mathsf{E}\varphi$ means φ is a path formulae satisfied on *at least one* possible future.

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

(Singleton g i) means g is a singleton group with i as the only member. (Agt α i) means i is the only agent of action α.

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

To represent an action α *achieving* a goal φ, we introduce a derived operator Achieves.

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

We will have a number of occasions to write A(Happens α), (action α occurs next in all alternative futures), and A¬(Happens α) (action α does not occur next in any alternative future), and so we introduce abbreviations for these.

$$(\text{Does } \alpha) \stackrel{\text{def}}{=} \mathsf{A}(\text{Happens } \alpha) \qquad (\text{Doesn't } \alpha) \stackrel{\text{def}}{=} \mathsf{A}\neg(\text{Happens } \alpha)$$

We find it convenient to define knowledge as true belief, rather than by introducing it as yet another primitive modality.

$$(\text{Know } i\ \varphi) \stackrel{\text{def}}{=} \varphi \wedge (\text{Bel } i\ \varphi)$$

We also find it convenient to use the notions of *mutual* mental states. Although we recognise that such states are idealised, in that they are not realisable in systems which admit the possibility of failed communication, they are nevertheless valuable abstraction tools for understanding multi-agent systems. The mutual belief of φ in a group of agents g is written (M-Bel g φ); the mutual goal of φ in g is written (M-Goal g φ), and the mutual knowledge of φ is written (M-Know g φ). We define mutual mental states as *fixed points*.

$$(\text{M-Bel } g\ \varphi) \stackrel{\text{def}}{=} \forall i \cdot (i \in g) \Rightarrow (\text{Bel } i\ \varphi \wedge (\text{M-Bel } g\ \varphi))$$
$$(\text{M-Goal } g\ \varphi) \stackrel{\text{def}}{=} \forall i \cdot (i \in g) \Rightarrow (\text{M-Bel } g\ (\text{Goal } i\ \varphi))$$
$$(\text{M-Know } g\ \varphi) \stackrel{\text{def}}{=} \varphi \wedge (\text{M-Bel } g\ (\text{M-Know } g\ \varphi))$$

3 Commitments, Conventions, and Intentions

The key mental states that control agent behaviour are intentions and joint intentions — the former define local asocial behaviour, the latter control social behaviour [2]. Intentions are important as they provide both the stability and predictability (through the notion of commitment) that is needed for social interactions, and the flexibility and reactivity (through the mechanisms by which commitments are monitored) that are required to deal with a changing environment. Previous attempts to formalise (joint) intentions have made no distinction between a commitment and its underlying 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 both the conditions under which a commitment might be abandoned, and how an agent should behave, should such a circumstance arise [8].

Commitments have a number of important properties (see [8] and [3, 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 [3].

When a group of agents are engaged in a cooperative activity, they have a joint commitment to the overall aim, as well as individual commitments to the specific tasks that they have been assigned. This joint commitment is parameterised by a social convention, which identifies the conditions under which the joint commitment can be dropped, and also describes how the agent should behave towards 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' inherent in joint action that it informs 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.

Formally, we define a convention as a set of rules, each rule consisting of a re-evaluation condition ρ and a goal γ: 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.

Definition 1. *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, $\forall k \in \{1, \ldots, l\}$.*

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.

> **Definition: (Joint commitments)** A group g is jointly committed to goal φ with respect to motivation ψ, pre-condition *pre*, and convention c iff: (i) pre-condition *pre* is initially satisfied; and (ii) until the termination condition is satisfied, every agent in g either (a) has a goal of φ; or (b) believes that the re-evaluation condition of some rule in c is satisfied, and has the goal corresponding to that re-evaluation condition; where the termination condition is that the goal part of some convention rule is satisfied.

More formally:

Definition 2. *If* $c = \{(\rho_k, \gamma_k) \mid k \in \{1, \ldots, l\}\}$ *is a convention, then:*

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

where

$$p \stackrel{\text{def}}{=} (\text{Goal } i\ \varphi) \qquad q \stackrel{\text{def}}{=} \bigvee_{l=1}^{k} (\text{Bel } i\ \rho_l) \wedge \mathsf{A}[(\text{Goal } i\ \gamma_l)\,\mathcal{U}\, r] \qquad r \stackrel{\text{def}}{=} \bigvee_{m=1}^{k} \gamma_m.$$

This general model can be used to capture the properties of many different types of joint commitment. For example, we will now specify a social convention that is similar to the Levesque-Cohen model of joint persistent goals (JPGs) [10]. Let

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

$$c_{JPG} \stackrel{\text{def}}{=} \left\{ \begin{array}{l} ((\text{Bel } i\ \varphi), (\text{M-Bel } g\ \varphi)), \\ ((\text{Bel } i\ \mathsf{A}\Box\neg\varphi), (\text{M-Bel } g\ \mathsf{A}\Box\neg\varphi)), \\ ((\text{Bel } i\ \neg\psi), (\text{M-Bel } g\ \neg\psi)) \end{array} \right\}.$$

A group with a joint commitment parameterised by a pre-condition pre_{JPG}, and convention c_{JPG} will have a shared mental state identical in all important respects to that implied by the JPGs of Levesque-Cohen. We use joint commitments to define joint intentions, which are held by a group g with respect to an action α and motivation ψ. In general, it is possible to make conventions a parameter of joint intentions. However, this would complicate our subsequent formalism, and we therefore leave this refinement to future work. For the purposes of this paper, we simply assume that joint intentions are defined over the JPG-like convention c_{JPG}; this gives us a model of joint intentions similar to that in [10, p98].

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

Thus a joint intention in g to do α means having a joint commitment that eventually g will believe α will happen next, and then α happens next. An individual intention by agent i to do α with motivation ψ is a special case of joint intention.

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

4 The Cooperative Problem Solving Process

In this section, we present a four-stage model of CPS, which we formalise by expressing it in the logic described in §2. The four stages of the model 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 is unable to achieve in isolation, or, more generally, because the agent prefers assistance.
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 having a joint commitment to collective action.
3. Plan formation: During this stage, the agents attempt to negotiate a joint plan that 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 an agreed social convention, which every agent follows.

Although we believe that most instances of CPS exhibit these stages in some form, we stress that the model is *idealised*. We recognise that there are cases which the model cannot account for, and we highlight these wherever appropriate. Our aim has been to construct a framework that describes CPS from beginning to end, but is *abstract* (in that details which might obscure more significant points have been omitted). (We once again stress that although space restrictions mean that we cannot completely define the logic used to represent the model here, a complete definition *is* presented in [17].)

4.1 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 its goal in isolation, due to a lack of resources, but believes that cooperative action can achieve it. For example, an agent may have a goal that, to achieve, requires information only accessible to another agent; without the cooperation of this other agent, the goal cannot be achieved.
- Alternatively, an agent may have the resources to achieve the goal, but does not want to use them. There may be several reasons for this: it may believe that in working alone on this particular problem, it will clobber one of its other goals, or it may believe that a cooperative solution will in some way be better (e.g., derived faster, more accurate).

In order to more 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- and multi-agent *ability*: what it means to be able to

bring about some state of the world. Several attempts to define multi-agent ability have appeared in the literature (e.g., [14]). However, there is currently no consensus on the appropriateness of these definitions. For this reason, we adapt the well-known model of ability proposed by Moore [12].

> **Definition: (Single-agent ability)** Agent i can achieve φ iff there is some possibly complex action α of which i is the sole agent, such that either: (i) i knows that after it performed α, φ would be satisfied; or (ii) i knows that after it performed α, it could achieve φ.

Clause (i) is the base case, where an agent knows the identity of an action that will achieve the goal φ directly. Clause (ii) allows for the possibility of an agent performing an action in order to find out how to achieve φ. This recursive definition is easily generalised to the multi-agent case.

> **Definition: (Multi-agent ability)** Group g can achieve φ iff there is some possibly complex action α and some group g', such that it is mutually known in g that $g' \subseteq g$, and g' are the agents of α, and it is mutually known in g that either (i) after α was performed, φ would be satisfied; or (ii) after α was performed, g would have the multi-agent ability to achieve φ.

Once again, clause (i) represents the base case, where the group is mutually aware of the identity of some action that could be performed by some subset of the group (whose identity must also be known), such that performing the action would achieve the goal directly. Clause (ii) is the recursive case, where the group is required to know the identity of some action and subset of agents such that performing the action would bring them closer to the goal.

A more precise definition of potential for cooperation can now be given.

> **Definition: (Potential for cooperation)** With respect to agent i's goal φ, there is potential for cooperation iff: (i) there is some group g such that i believes that g can jointly achieve φ; and either (ii) i can't achieve φ in isolation; or (iii) i believes that for every action α that it could perform which achieves φ, it has a goal of not performing α.

Note that in clause (i), an agent needs to know the identity of a group that it believes can cooperate to achieve its goal. This is an overstrong 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 CNET [15]). However, catering for these cases would complicate the formalisation a good deal, and obscure some more important points. We therefore leave such refinements to future work.

The ideas introduced above are readily expressed using the language we described in §2. First, we write $(\mathsf{Can}\ i\ \varphi)$ iff i can achieve φ in isolation.

$$\begin{aligned}(\mathsf{Can}\ i\ \varphi) \stackrel{\text{def}}{=}\ & \exists \alpha \cdot (\mathsf{Know}\ i\ (\mathsf{Agt}\ \alpha\ i) \wedge (\mathsf{Achieves}\ \alpha\ \varphi)) \quad \vee \\ & \exists \alpha \cdot (\mathsf{Know}\ i\ (\mathsf{Agt}\ \alpha\ i) \wedge (\mathsf{Achieves}\ \alpha\ (\mathsf{Can}\ i\ \varphi)))\end{aligned}$$

Multi-agent ability is a generalisation of single-agent ability.

$$(\text{J-Can } g\ \varphi) \stackrel{\text{def}}{=} \exists\alpha \cdot \exists g' \cdot (\text{M-Know } g\ (g' \subseteq g) \wedge (\text{Agts } \alpha\ g') \wedge (\text{Achieves } \alpha\ \varphi)) \quad \vee$$
$$\exists\alpha \cdot \exists g' \cdot (\text{M-Know } g\ (g' \subseteq g) \wedge (\text{Agts } \alpha\ g') \wedge (\text{Achieves } \alpha\ (\text{J-Can } g\ \varphi)))$$

We can now formally state the conditions that characterise the potential for cooperation.

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

4.2 Team Formation

Having identified the potential for cooperative action with respect to one of its goals, a rational agent will solicit assistance from some 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 a mental state wherein the group has a joint commitment to collective action. (There will not yet be a joint intention to act; this comes later.) An agent cannot guarantee that it will be successful in forming a team; it can only *attempt* it. We adapt the model of attempts developed by Cohen-Levesque [4, p240].

> **Definition: (Attempts)** An attempt by agent i to bring about a state φ is an action α performed by i with the goal that after α is performed, φ is satisfied, or 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' [4, p240]. If i is successful, then bringing about ψ will be sufficient to cause φ.

The team formation stage can then be characterised as an assumption made about rational agents: namely, that an agent which recognises the potential for cooperative action will solicit assistance.

> **Assumption: (Team formation)** 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, (that it believes can jointly achieve φ), a state wherein: (i) it is mutually believed in g that g can jointly achieve φ, and g are jointly committed to team action with respect to i's goal φ; or, failing that, to at least cause in g (ii) the mutual belief that i has a goal of φ and the mutual belief that i believes g can jointly achieve φ.

Part (i) represents the commitment that the group has towards i's goal φ if i is successful in its attempt to solicit assistance; we discuss what team action means in §4.4. Note that an agent might have its own reasons for agreeing to participate in a cooperative action, that are unconnected with the request by the agent that recognises the potential for cooperation. However, we have not attempted to deal with such cases here.

The team formation assumption implicitly states 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. (We refer the interested reader to [7, pp159–165] for a discussion and formalisation of these considerations.)

We write $\{\mathsf{Attempt}\ i\ \alpha\ \varphi\ \psi\}$ for an attempt by i to achieve φ by performing α, at least achieving ψ. Following Cohen-Levesque, we use curly brackets to indicate that attempts are complex actions, not predicates [4, p240].

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

We introduce an abbreviation to simplify subsequent formalisation: (Pre-Team $g\ \varphi\ i$) means that (i) g mutually believe that they can jointly achieve φ; and (ii) g are jointly committed to becoming a team with respect to i's goal φ.

$$\begin{array}{l}(\mathsf{Pre\text{-}Team}\ g\ \varphi\ i) \stackrel{\text{def}}{=} (\mathsf{M\text{-}Bel}\ g\ (\mathsf{J\text{-}Can}\ g\ \varphi)) \wedge \\ \qquad\qquad (\mathsf{J\text{-}Commit}\ g\ (\mathsf{Team}\ g\ \varphi\ i)\ (\mathsf{Goal}\ i\ \varphi)\ pre_{JPG}\ c_{JPG})\end{array}$$

(Team is defined in §4.4.) The main assumption concerning team formation can now be stated.

Assumption 1 $\models \forall i \cdot (\mathsf{Bel}\ i\ (\mathsf{PfC}\ i\ \varphi)) \Rightarrow \mathsf{A}\Diamond\exists g \cdot \exists\alpha \cdot (\mathsf{Happens}\ \{\mathsf{Attempt}\ i\ \alpha\ p\ q\})$ *where*

$$\begin{array}{l}p \stackrel{\text{def}}{=} (\mathsf{Pre\text{-}Team}\ g\ \varphi\ i) \\ q \stackrel{\text{def}}{=} (\mathsf{M\text{-}Bel}\ g\ (\mathsf{Goal}\ i\ \varphi) \wedge (\mathsf{Bel}\ i\ (\mathsf{J\text{-}Can}\ g\ \varphi))).\end{array}$$

If team formation is successful then for the first time there will be a social mental state relating to i's goal, which contrasts with i's individual perspective that has guided the process until this stage.

4.3 Plan Formation

If an agent is successful in its attempt to solicit assistance, then there will be a group of agents with a joint commitment to collective action. But collective action cannot 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 that there is at least one action that is known to the group that will take them 'closer' to the goal (see the definition of J-Can, above). However, it is possible that there are many agents that know of actions the group can perform in order to take the collective closer to, or even achieve the goal. Moreover, some members of the collective may have objections to one or more of these

actions. For example, an agent may believe that a particular action has hitherto unforeseen and damaging consequences. It is therefore necessary for the collective to come to some agreement about exactly which course of action they will follow. *Negotiation* is the mechanism via which such agreement is reached.

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 reached agreement[2]. Negotiation has long been recognised as a process of some importance for DAI (see, e.g., [16]). Unfortunately, analyses of negotiation demonstrate that it is also extremely complex — a rigorous attempt at formalisation is quite beyond the scope of this paper[3]. 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 that 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 — we shall say no more about team action here, as this is the subject of the next section.

We can make a number of other tentative assumptions about the behaviour of agents during negotiation. Most importantly, we might assume that they 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.

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.

> **Definition: (Joint 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).

Next, we state the minimum conditions required for negotiation to have occurred.

> **Assumption: (Negotiation)** 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 where it is mutually known in g that 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.

In other words, 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

[2] It may also involve agents lying, or being cunning and devious, though we shall not consider such cases here.

[3] But see [9] for preliminary work on logical models of argumentation.

at least one of them has proposed a plan that it believed would achieve the desired goal. The other, more tentative assumptions about agent behaviour during negotiation are as follows.

> **Assumption: (Making preferences known)** 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.

We are once again assuming that agents are veracious, in that they attempt to influence the team by revealing their true preferences, rather than by lying, or concealing their true preferences.

We begin by formalising joint attempts.

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

The main assumption characterising negotiation can now be given. (Team is defined below.)

Assumption 2 $\models (\text{Pre-Team } g\ \varphi\ i) \Rightarrow \mathsf{A}\Diamond\exists\alpha \cdot (\text{Happens } \{\text{J-Attempt } g\ \alpha\ p\ q\})$ *where*

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

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.

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

The definition of (Try-to-prevent $i\ g\ \alpha$) is similar to Try-to-cause, and is therefore omitted.

Assumption 3 *Agents who have a preference for some action make the team mutually aware of their preference:*

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

Agents who prefer some action not to be performed make the team mutually aware of their preference:

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

If plan formation is successful then the team will have a joint commitment to the goal, and will have agreed to the means by which they will pursue this goal. 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 (see §5).

4.4 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 jointly intend some appropriate action.

> **Definition: (Team action)** A group g are considered a team with respect to i's goal φ iff there is some action α, such that: (i) α achieves φ; and (ii) g have a joint intention of α, relative to i having a goal of φ.

The formalisation of Team is simple.

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

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 the social convention dictates that the agent will make the team aware of this, and team action will end.

5 Concluding Remarks

In this paper, we have presented an abstract formal model of cooperative problem solving, which describes all aspects of the process, from recognition of the potential for cooperation through to team action. This model considers a number of issues that have hitherto been neglected by DAI theorists. For example, it defines the conditions under which there is potential for cooperative action, and shows how an agent's individual mental state can lead it to attempt to build a social mental state in a group. The model has a number of other properties, which we shall briefly discuss in this section.

Although we have not explicitly considered communication, our model is nevertheless consistent with one of the best current theories of *speech acts*: in [4], Cohen-Levesque proposed a theory in which illocutionary acts are treated as *attempts* to bring about some mental state in a conversation participant. At a number of points, our model predicts precisely such attempts; for example, the model predicts that an agent which recognises the potential for cooperation will attempt to bring about a joint commitment to collective action in some group that it believes can achieve its goal.

Another interesting property is that the model consists of a set of *liveness* properties [11]. 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.

The model also predicts that agents will attempt to initiate social interaction if they have goals that are dependent on other community members. In order to do this, the agents must have some knowledge about the abilities, skills, and interests of their acquaintances.

Finally, the model predicts that once a group of 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 proposals; they are thus autonomous, rather than benevolent.

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, as highlighted in the main text. Additionally, there are a number of ways in which the language we have used for representing the model needs to be extended. The two most significant points are the need to quantify over complex action expressions, and the need to be able to represent meta-level notions at the object level.

References

1. A. H. Bond and L. Gasser, editors. *Readings in Distributed Artificial Intelligence*. Morgan Kaufmann Publishers: San Mateo, CA, 1988.
2. M. E. Bratman. *Intentions, Plans, and Practical Reason*. Harvard University Press: Cambridge, MA, 1987.
3. P. R. Cohen and H. J. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42:213–261, 1990.
4. 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.
5. E. H. Durfee. *Coordination of Distributed Problem Solvers*. Kluwer Academic Publishers: Boston, MA, 1988.
6. 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.
7. J. R. Galliers. *A Theoretical Framework for Computer Models of Cooperative Dialogue, Acknowledging Multi-Agent Conflict*. PhD thesis, Open University, UK, 1988.
8. N. R. Jennings. Commitments and conventions: The foundation of coordination in multi-agent systems. *Knowledge Engineering Review*, 8(3):223–250, 1993.
9. S. Kraus, M. Nirke, and K. Sycara. Reaching agreements through argumentation: A logical model. In *Proceedings of the Twelfth International Workshop on Distributed Artificial Intelligence (IWDAI-93)*, pages 233–247, Hidden Valley, PA, May 1993.
10. 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.
11. Z. Manna and A. Pnueli. *The Temporal Logic of Reactive and Concurrent Systems*. Springer-Verlag: Heidelberg, Germany, 1992.

12. 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.
13. A. S. Rao and M. P. Georgeff. Social plans: Preliminary report. In E. Werner and Y. Demazeau, editors, *Decentralized AI 3 — Proceedings of the Third European Workshop on Modelling Autonomous Agents and Multi-Agent Worlds (MAAMAW-91)*, pages 57–76. Elsevier Science Publishers B.V.: Amsterdam, The Netherlands, 1992.
14. M. P. Singh. Group ability and structure. In Y. Demazeau and J.-P. Müller, editors, *Decentralized AI 2 — Proceedings of the Second European Workshop on Modelling Autonomous Agents and Multi-Agent Worlds (MAAMAW-90)*, pages 127–146. Elsevier Science Publishers B.V.: Amsterdam, The Netherlands, 1991.
15. R. G. Smith. *A Framework for Distributed Problem Solving*. UMI Research Press, 1980.
16. 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.
17. M. Wooldridge. Coherent social action. In *Proceedings of the Eleventh European Conference on Artificial Intelligence (ECAI-94)*, pages 279–283, Amsterdam, The Netherlands, 1994.