ON BEING RESPONSIBLE[1]

N.R.Jennings
Dept Electronic Engineering,
Queen Mary & Westfield College,
London E1 4NS
UK.
n.r.jennings@qmw.ac.uk

Joint responsibility is a mental and behavioural state which captures and formalizes many of the intuitive underpinnings of collaborative problem solving. It defines the pre-conditions which must hold before such activity can commence, how individuals should behave (in their own problem solving and towards others) once such problem solving has begun and minimum conditions which group participants must satisfy.

1. JOINT ACTION: AN INTRODUCTION

In an environment composed of multiple agents there are many different forms of social activity which can occur (eg cooperation, competition and hostility). The aim of this paper is to provide a framework in which one particular class of social activity can be formalised and ultimately analysed: namely that in which a group of autonomous agents (at least two) decides they wish to work together as a team to solve a common problem. A comprehensive theory describing this class of social interaction would need to cover at least the following aspects: when to initiate team activity, how to go about assembling the team, how to plan and distribute work within the team, how to behave once team activity has been initiated and how to complete team activity. The framework described herein defines the prerequisites for such action and also prescribes how agents should behave (both in their own problem solving and with respect to other group members) once the problem solving has been established.

Typically in a community of autonomous agents, one of the primary motives for joint action is when no individual is capable of achieving a desired objective alone; only by combining and coordinating with others can the target be reached. Joint action is usually a reciprocal process in which participating agents augment their objectives and problem solving to comply with those of others - hence it is a fairly sophisticated form of cooperation. It requires greater knowledge, awareness and reflection by an agent both with respect to its own problem solving objectives and about their compatibility with the objectives of others, than simpler forms of social interaction (such as task and result sharing [19]).

Joint action, by definition, requires an objective the group wishes to achieve - it is the glue which binds the team together. As a consequence of the autonomous nature of the agents, team members will only participate if they can derive some benefit from the interaction (i.e. benevo-

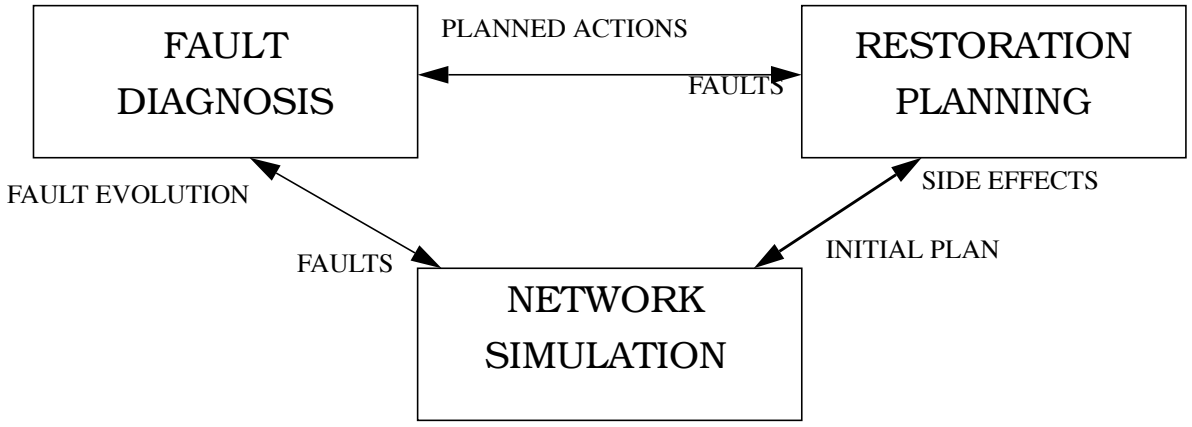
[1]. The work described in this paper has been partially supported by the ESPRIT II project P2256 (ARCHON) whose partners are: Krupp Atlas Elektronik, Ispra, Framentec, Labein, QMW, IRIDIA, Iberduero, ERDC, Amber, University of Athens, Univ. of Amsterdam, Volmac, CERN and Univ. of Porto.

lence is not assumed [6], [23]). However merely having a common objective is not sufficient for realising a collective goal - agents also need to agree upon a means of reaching the target state. Previous work on collaborative problem solving [11], [16], [17], [18], [20] has concentrated on defining joint intentionality in terms of goal states, without consideration of how goals can be achieved (plan states) or how participants should behave when engaged in collaborative problem solving. The fundamental notion behind our work is that any comprehensive description of joint action should include all of these aspects. Therefore we present a mental and behavioural state (called *joint responsibility*) which internalises for each team member the notion of solving problems in a group. Joint responsibility defines the pre-conditions necessary for joint action as well as descriptions of how others will act, both in performing their social problem solving and in participating in collaborative interaction per se. This formalization is especially important if agents are to operate in dynamic and complex environments in which their aims and objectives are likely to alter during the course of a prolonged cooperative interaction. Also if we are to move to environments in which agents are more autonomous in nature, then it is important that many of the assumptions presently hidden in the agents' control structures are made explicit and can be subjected to an agents reasoning processes.

In the remainder of this paper the notion of joint responsibility as a pre-requisite for joint action is developed. Section two describes a scenario in which joint problem solving by a team of autonomous agents is beneficial to all participants and also enhances the quality of the output to the user, while section three introduces and formalizes the notion of joint responsibility.

2. FAULT RECOVERY IN ELECTRICITY DISTRIBUTION NETWORKS

The domain in which the principles related to joint action will be illustrated is that of fault recovery in electricity distribution networks - this example is loosely based on an ARCHON application [12], [13], [21]. The scenario involves three pre-existing systems each of which has a set of clearly defined goals and is capable of sophisticated problem solving in its own right.



The fault diagnosis agent can detect and indicate to the other two that a fault has occurred in the network. The restoration planning agent (RA) is responsible for constructing a maintenance plan once a fault has been detected - such a plan will instruct the operator to perform certain sequences of operations in a well defined order. The network simulation agent (NSA) is capable of running and explaining "what-if" simulations of the network based on the settings of certain key parameters. One joint action which can be instantiated between the NSA and RA is in the area of producing restoration plans whose actions will not cause further parts of the

network to fail (i.e. “sensible” restoration plans). As a standalone system, the plans suggested by the RA may lead to further faults as the specified operations can overload currently working components. However if the RA’s tentative restoration plan is first sent to the NSA then its effects can be predicted, problem areas highlighted and the plan refined accordingly. If major problems are identified and the RA decides to significantly revise its restoration plan then there may be further interaction with the NSA. If only minor modifications are made or the RA deems the highlighted risks acceptable, then no further interaction will be necessary.

3. JOINT RESPONSIBILITY

Joint responsibility defines the conditions which need to be satisfied before joint action can be initiated and a code of conduct specifying how agents should react when the joint action becomes unsustainable. Joint responsibility will be defined using a logical formalism, similar to that described in [4]. This formalism has the usual connectives of a first order language (^ AND, v OR, ~NOT) - as well as operators for propositional attitudes. BEL(x, p) and GOAL(x, p) mean agent x has p as a belief and a goal respectively, MB({x, y}, p) that x and y mutually believe p. Dynamic logic constructs are also used [8]: ❑p means p is always true and ◊p that p will eventually be true. p?;a means “action a with p holding initially”, and analogously for a;p?. This analysis will be exemplified using the restoration planning agent (RA) and the network simulation agent (NSA), although it can of course be applied to groups of arbitrary size.

3.1 Common Goals and Joint Persistence

The first step to achieving joint action is that a group of two or more agents realize that they have a common objective (*intention*[2]) and that this can only (best) be fulfilled by collaborating with others. Once this is believed by all participants, a common goal exists and each individual becomes committed [4] to achieving that objective. However as Levesque et al. point out, this is not a sufficiently sturdy foundation upon which robust *joint* action can be based [11]; it is particularly fragile if agents intentions change (i.e. they reach a state in which they are no longer committed to attaining the common objective). To rectify these problems, they propose the notion of joint persistent goals (JPGs) [11] in which groups of agents become jointly committed to a common aim. The properties of JPGs can best be illustrated using an example. Suppose the RA and the NSA have established a JPG of producing a sensible restoration plan and then at some later stage one of the agents (say the NSA) no longer desires this objective (because the user has asked it to run a what-if question on the network as a high priority task). Should it simply drop the common goal without informing the RA?, meaning that the RA will be left waiting indefinitely. Clearly not! Therefore in the interests of robust group problem solving, a JPG requires that the NSA adopts the goal of informing the RA of its change of intention. Thus, JPGs define the conditions under which a commitment to a joint goal can be dropped and also how participants should act when they find themselves in such a situation.

3.2 Solution Commitment

Contrary to the claims of Levesque et al. [11], having a JPG is not sufficient for obtaining joint action. JPG’s only specify that agents have a common desire to reach a target state, they do not specify *how* to reach this state. Agreeing upon a means of reaching the state is nearly as impor-

[2] Intentions have been ascribed a variety of differing meanings (eg [3], [4], [24])- within this context they specify a desired or target state, *without* consideration of how that state is to be attained.

tant as the desire to reach the state itself. Therefore although NSA and RA may be able to agree that they want to produce a restoration plan together, unless they can agree upon a common means of achieving this then joint action will not follow. In some circumstances, such agreement may be impossible because of the autonomous nature of the agents involved; both agents may have several objectives at any one time and these must be balanced with the desire to produce a sensible restoration plan. If they have insufficient resources or it conflicts with other more important intentions then it may be impossible for them to converge upon a common solution, even though they share a common objective.

There are several facets to agreeing upon a common solution, firstly the strategy by which the solution will be produced (i.e. the group's organizational structure) and then development and agreement of the actual plan steps. Within this framework we are not concerned with the mechanisms used for achieving the common solution; rather we are concerned with the fact that they must agree upon the *principle* that a common plan is needed to tackle the joint problem.

3.2.1 Plan Representation Language

In a multi-agent environment the plan representation language must include the agent(s) which will perform the action as well as the action itself. In situations in which agents act collaboratively it is essential to be able to describe activities in terms of groups of agents, rather than just in terms of individuals[3]. Therefore the fact that a group of agents $\{\alpha_1,..,\alpha_n\}$ will work together in order to try and achieve σ will be represented as follows: $< \{\alpha_1,..,\alpha_n\}, \sigma >$. Let the set of all agents existing in the environment be denoted by A; unless stated to the contrary, all groups of agents are a subset of the members of A.

Intentions will typically be composed of sub-intentions which are themselves decomposable - the solution graph for σ being represented by Σ_σ. The nodes without successors (when the graph has been fully expanded) correspond to atomic units of activity (*primitive actions*) which are executable by individual agents. The various stages of intention execution[4] can be expressed as follows:

$$\text{EXECUTE/EXECUTING/EXECUTED}(< \{\alpha_1,..,\alpha_n\}, \sigma >, \Sigma_\sigma)$$

which respectively mean that Σ_σ will be executed next, is being executed now or has been executed, for the purpose of achieving σ by agents $\{\alpha_1,..\alpha_n\}$. Underlying this definition is the assumption that at least one team member (or a subset of them) is (are) capable of realising the constituent sub-intentions and that team members will not attempt actions which they cannot execute to some degree.

3.2.2 Inter-Related Actions

Typically the solution of a joint action contains some actions which need to be coordinated with those of other agents and some which can be executed independently. Solutions will therefore contain interrelated components: $\Sigma_\sigma = \{\sigma_1\ \Re_{1,2}\ \sigma_2,\ \sigma_3\ \Re_{3,4}\ \sigma_4,....\}$. $\Re_{1,2}$ defines the relationship between σ_1 and σ_2[5] (see Allen's taxonomy of temporal relationships [1] or

3. This representation is consistent with social psychology work in which the society is considered prior to the individual, not the other way around [15] and has also been independently noted in [10] & [16].

4. Execution in this context corresponds to searching through the space of partial plans if a node is expandable or processing a primitive action if not.

5. $\Re$ is a non-commutative, non-associative n-ary operation ($n \geq 2$).

plan transformations of action-ordering plans [9]). Such relationships are an integral component of the solution specification and if they are not satisfied then the desired objective cannot be guaranteed by solution Σ_σ. Hence fulfilling an intention means performing the actions and satisfying any relationships which exist with other actions[6]:

$$(\forall <\{\alpha_w... \alpha_x\}, \sigma_i> \in \Sigma_\sigma) (\exists <\{\alpha_y... \alpha_z\}, \sigma_j> \in \Sigma_\sigma) \Re_{i,j} \supset$$
$$\text{MB}(\{\alpha_w... \alpha_x\}, \Re_{i,j})?; \text{EXECUTE}(<\{\alpha_w... \alpha_x\}, \sigma_i>, \Sigma_{\sigma i})$$

The success of a solution in reaching its desired objective is the final component of the plan representation language:

$$\text{ACHIEVE}(<\{\alpha_1,..,\alpha_n\}, \sigma>, \Sigma_\sigma) \Leftrightarrow \sim\sigma \wedge \text{EXECUTE}(<\{\alpha_1,..,\alpha_n\}, \sigma>, \Sigma_\sigma); \sigma?$$

meaning if Σ_σ is executed next, σ which did not hold before this sequence of actions, will hold as a direct consequence of performing the specified actions.

3.2.3 Example of Formalism

To illustrate this formalism, the intention of producing a sensible restoration plan can be expressed as follows: σ = < {RA, NSA}, SENSIBLE-RESTORATION-PLAN> and one solution for achieving this is:

Σ_σ = {<RA, TENTATIVE-RESTORATION-PLAN> ***BEFORE***
<NSA, CHECK-PLAN-FOR-OVERLOADS>

<NSA, CHECK-PLAN-FOR-OVERLOADS> ***BEFORE***
<RA, REFINE-RESTORATION-PLAN>}

3.2.4 Defining Solution Commitment

It is now possible to express the second pre-condition for joint action (the first being the existence of a common objective), namely: that the participants must agree upon the *principle* that a common solution is needed if the objective is to be achieved:

$$\text{NEED-COMMON-SOLUTION} (<\{\alpha_1,..,\alpha_n\}, \sigma>) \Leftrightarrow$$
$$\text{MB}(\{\alpha_1,..,\alpha_n\}, \Diamond \exists\Sigma_\sigma \text{EXECUTE}(<\{\alpha_1,..,\alpha_n\}, \sigma>, \Sigma_\sigma) \vee \Box\sim\sigma)$$

The idea that joint action requires a common solution is also expressed either implicitly (eg through the definition of roles [24]) or explicitly (eg [7], [14]) in most joint intention work. However what is lacking in previous work is a prescription of how team members should behave once such a solution has been developed. The notion of solution commitment fills this gap: defining under what conditions an agent should and should not follow the agreed solution and how it should behave when it is no longer rational for it to keep to the solution. Before the complete set of circumstances can be described the terms invalid, unattainable and violated need to be defined (illustrations are taken from the sensible restoration plan example):

Definition: *Invalid Plan* - following the plan no longer leads to the desired goal

$$\text{INVALID} (<\{\alpha_1,..\alpha_n\}, \sigma>, \Sigma_\sigma) \Leftrightarrow \Box\sim\text{ACHIEVE}(<\{\alpha_1,..,\alpha_n\}, \sigma>, \Sigma_\sigma)$$

6. Fulfilling relationships is a process requiring communication and synchronization between the responsible agents. Details of how this is achieved are not described in this paper

eg the NSA believes the network status has changed substantially since the simulation to judge the effect of the tentative restoration plan was started; meaning its analysis will be inaccurate as essential information is missing. Hence it is impossible to tell whether the restoration plan is safe or not without redoing the simulation - a futile activity because the tentative restoration plan will be significantly altered by the new information. Therefore following the agreed solution will not produce the desired result.

Definition: *Unattainable Plan* - one of the specified plan steps cannot be executed

$$\text{UNATTAINABLE}(<\{\alpha_1,..\alpha_n\}, \sigma>, \Sigma_\sigma) \Leftrightarrow$$
$$(\exists <\{\alpha_w,..,\alpha_x\}, \sigma_i> \in \Sigma_\sigma) \text{❑}\sim\text{EXECUTE}(<\{\alpha_w,..,\alpha_x\}, \sigma_i> \Sigma_\sigma)$$
$$\text{where } [\{\alpha_w,..,\alpha_x\} \subseteq \{\alpha_1,..,\alpha_n\}]$$

eg to simulate the effects of the RA's restoration plan, the NSA has to set simulation parameters outside their permitted range - meaning the simulation cannot be executed.

Definition: *Violated Plan* - one of the plan steps which should have been performed has not been performed, or a relationship between plan steps has not been upheld. This differs from unattainability in that the action could feasibly have been executed, but the agents involved did not do so.

$$\text{VIOLATED}(<\{\alpha_1,..\alpha_n\}, \sigma>, \Sigma_\sigma) \Leftrightarrow \sim\text{EXECUTED}(<\{\alpha_1,..,\alpha_n\}, \sigma>, \Sigma_\sigma)$$

eg the RA has sent a tentative plan to the NSA which has run a simulation and highlighted potential problem areas. However because a major incident has occurred on the network, the RA decides that rather than refining the tentative plan it is better to try and generate a new plan because the information will be more up-to-date. In this case, the RA has violated the agreed plan by not performing the refinement task.

In addition to these plan states, it would be irrational for an agent to remain committed to a solution if either the plan's objective already holds or another team member is no longer committed to the agreed solution. An example of the former situation is if the NSA calculates that the proposed tentative plan will cause no additional problems in the network - therefore the joint objective of producing a sensible restoration plan has already been met and no additional work is required. An example of the latter occurs if the RA no longer believes it possible to produce a restoration plan given the current context, then there is no point in the NSA evaluating the proposed restoration plan as the result will simply be irrelevant to the RA.

We are now in a position to define the conditions under which it is rational for an agent to stop performing actions specified in the agreed solution. Hence, unless these conditions prevail, an individual agent (α) should remain committed (I-COMMIT-CONDS) to solution Σ_σ as a means of achieving σ:

$$\text{I-COMMIT-CONDS}(\alpha, <\{\alpha_1,..\alpha_n\}, \sigma>, \Sigma_\sigma) \Leftrightarrow \quad [\alpha \in \{\alpha_1,..,\alpha_n\}]$$
$$\text{BEL}(\alpha, \sim\sigma) \wedge$$
$$\text{BEL}(\alpha, \sim\text{INVALID}(<\{\alpha_1,..\alpha_n\}, \sigma>, \Sigma_\sigma)) \wedge$$
$$\text{BEL}(\alpha, \sim\text{UNATTAINABLE}(<\{\alpha_1,..\alpha_n\}, \sigma>, \Sigma_\sigma)) \wedge$$
$$\text{BEL}(\alpha, \sim\text{VIOLATED}(<\{\alpha_1,..\alpha_n\}, \sigma>, \Sigma_\sigma)) \wedge$$
$$\text{BEL}(\alpha, (\forall \alpha_i \in \{\alpha_1,..,\alpha_n\} \text{ I-COMMIT-CONDS}(\alpha_i, <\{\alpha_1,..\alpha_n\}, \sigma>, \Sigma_\sigma)))$$

Once an agent believes one (or more) of the above conditions to be true, it should stop performing all associated group problem solving activity and re-evaluate the situation. However merely stopping local activity fails to capture our pre-theoretic intuitions about what it means to solve problems in a group. In such circumstances, individuals must endeavour to inform other team members that they are no longer committed to the solution and the reason for this. Based on this intuition, it is now possible to formalize how individuals within the team should act once a common solution has been derived and agreed upon. This behaviour is called *individual solution commitment* (ISC):

$$\begin{aligned}
&\text{ISC}(\alpha, <\{\alpha_1,..\alpha_n\}, \sigma>, \Sigma_\sigma) \Leftrightarrow && [\alpha \in \{\alpha_1,..,\alpha_n\}] \\
&\quad \text{UNTIL } \sim\text{I-COMMIT-CONDS}(\alpha, <\{\alpha_1,..\alpha_n\}, \sigma>, \Sigma_\sigma), \\
&\quad\quad (\forall <\{\alpha_w,..\alpha_x\}, \sigma_i> \in \Sigma_\sigma) \wedge (\alpha \in \{\alpha_w,..\alpha_x\}) \supset && [\{\alpha_w...\alpha_x\} \subseteq \{\alpha_1...\alpha_n\}] \\
&\quad\quad\quad \text{BEL}(\alpha, \Diamond\text{EXECUTE}(<\{\alpha_w,..\alpha_x\}, \sigma_i>, \Sigma_{\sigma i})) \wedge \\
&\quad\quad\quad \text{EXECUTE}(<\{\alpha_w,..\alpha_x\}, \sigma_i>, \Sigma_{\sigma i})) \\
&\quad \text{WHEN GOAL}(\alpha, \text{MB}(\{\alpha_1,..\alpha_n\}, \sim\text{I-COMMIT-CONDS}(\alpha, <\{\alpha_1,..\alpha_n\}, \sigma>, \Sigma_\sigma)))^{7}
\end{aligned}$$

From this definition it is apparent that a group member will try and fulfil its obligations specified in the agreed solution whilst it is still committed to that solution as a means of achieving the desired result - on becoming uncommitted it endeavours to inform others of this fact. What happens when all team members are aware of the lack of ISC by an agent within their ranks depends upon the reason for the loss of commitment. Such behaviour can include stopping all associated activity indefinitely, replanning using existing team members or trying to recruit new members to the group - however the link between such actions and the reason for the of loss of ISC is left for another time.

Definition: Combining the results of this section, there are two facets concerned with actions for achieving a target state: there is the principle of agreeing to the need for a common solution and also a definition of how group members should behave once such a solution has been agreed upon. These two components can be joined together into a single proposition called *solution commitment*:

$$\begin{aligned}
&\text{SOLUTION-COMMITMENT}(<\{\alpha_1,..,\alpha_n\}, \sigma>) \Leftrightarrow \\
&\quad \text{MB}(\{\alpha_1,..,\alpha_n\}, \text{NEED-COMMON-SOLUTION}(<\{\alpha_1,..,\alpha_n\}, \sigma>)) \wedge \\
&\quad (\forall\ \alpha_i \in \{\alpha_1,..\alpha_n\}\ \text{ISC}(\alpha_i, <\{\alpha_1,..\alpha_n\}, \sigma>, \Sigma_\sigma))
\end{aligned}$$

Solution commitment has been developed independently of the group organization structure and mechanism used for reaching the common solution; therefore it is applicable across a wide range of paradigms - one agent planning for all others [22], the plan being developed collaboratively [5], being developed in one go or through incremental refinement, decisions taken by majority or requiring consensus and so on. Returning to our example, this means the RA and the NSA have to agree upon the principle that a common plan is needed for producing a sensible restoration plan. Once agreed, a common solution can be developed and both agents will endeavour to do their parts (eg the RA will generate a tentative plan, the NSA will highlight any potential problem areas and the RA will then refine it based on this information). They will continue to do this until the task is completed satisfactorily, one of them finds the agreed solution unsustainable or discovers the other is no longer committed to the solution.

---

7. UNTIL p,q WHEN r: until p is true, q will remain true. When (if) p becomes true, r will become true

3.3 Contributions

An important attribute which is missing from previous descriptions of joint action is that of group minimality; stated simply, to be included in a group an agent must be able to contribute something! In the restoration plan example, the joint action is between the RA and the NSA and it makes no sense for any other agent to be involved, because it would be unable to carry out useful problem solving in the cooperation context. This property of group members is both conceptual ("free-loading" in a group is undesirable) and pragmatic (the time and communication resource consumed coordinating group activity is usually proportional to the size of the group, therefore it makes sense to only include individuals who carry out activities beneficial to the group's objectives). There are two ways in which an agent can contribute to the attainment of a group goal: it can perform an act which is part of the agreed solution (*positive contribution*) or it may refrain from performing an action which would interfere with the agreed solution (*non-negative contribution*). Imagine a team of agents trying to stack blocks B1, B2 and B3 - a positive contribution could be putting B2 onto B1, a non-negative one not unstacking B2. However due to space limitations we will only consider positive contributions.

Definition: A sub-intention can potentially contribute to a parent intention if it is a component of any solution (however inefficient or cumbersome) which achieves the parent intention:

$$\text{CONTRIBUTES}(\langle\{\alpha_j,..,\alpha_k\}, \sigma_i\rangle, \langle\{\alpha_1,..,\alpha_n\}, \sigma\rangle) \Leftrightarrow$$
$$(\exists\, \Sigma_\sigma\ \text{ACHIEVE}(\langle\{\alpha_w,..,\alpha_x\}, \sigma\rangle, \Sigma_\sigma) \wedge \langle\{\alpha_j,..,\alpha_k\}, \sigma_i\rangle \in \Sigma_\sigma$$
$$\text{where } [\{\alpha_w,..,\alpha_x\} \subseteq \{\alpha_1,..,\alpha_n\}] \text{ and } [\{\alpha_j,..,\alpha_k\} \subseteq \{\alpha_w,..,\alpha_x\}]$$

The first stage is for the individual to believe that it is capable of offering something to the group. Once an individual is sure of this, it then has to convince others that it's inclusion will benefit the group. Concentrating on the former, an agent is capable of contributing to the group goal if it can achieve a sub-intention which is a component of a *potential* overall solution:

$$\text{CAN-CONTRIBUTE}(\alpha, \langle\{\alpha_1,..,\alpha_n\}, \sigma\rangle) \Leftrightarrow$$
$$\text{ACHIEVE}(\langle\{\alpha_w,..,\alpha_x\}, \sigma_i\rangle, \Sigma_{\sigma i}) \wedge \quad [\{\alpha_w,.,\alpha_x\} \subseteq \{\alpha_1,\alpha_n\}]$$
$$\text{CONTRIBUTES}(\langle\{\alpha_w,..,\alpha_x\}, \sigma_i\rangle, \langle\{\alpha_1,..,\alpha_n\}, \sigma\rangle) \quad [\alpha \in \{\alpha_w,..\alpha_x\}]$$

The potential of being able to contribute to the attainment of a goal is not sufficient to guarantee entry into a group; it must be believed that the individual *actually* intends to participate:

$$\text{WILL-PARTICIPATE}(\alpha, \langle\{\alpha_1,..,\alpha_n\}, \sigma\rangle) \Leftrightarrow \quad [\alpha \in \{\alpha_1,..\alpha_n\}]$$
$$\Diamond\text{SOLUTION-COMMITMENT}(\langle\{\alpha_1,..,\alpha_n\}, \sigma\rangle)$$

An agent will only be admitted into joint problem solving activity if all group members are firstly convinced that the agent is capable of contributing to the objective and secondly that they believe it will actually participate:

$$\text{MAY-CONTRIBUTE}(\alpha, \langle\{\alpha_1,..,\alpha_n\}, \sigma\rangle) \Leftrightarrow \quad [\alpha \in \{\alpha_1,..\alpha_n\}]$$
$$\text{MB}(\{\alpha_1,..,\alpha_n\}, \text{CAN-CONTRIBUTE}(\alpha, \langle\{\alpha_1,..,\alpha_n\}, \sigma\rangle)) \wedge$$
$$\text{MB}(\{\alpha_1,..,\alpha_n\}, \text{WILL-PARTICIPATE}(\alpha, \langle\{\alpha_1,..,\alpha_n\}, \sigma\rangle))$$

This is a conservative approach to setting up groups of cooperating agents, in that all team members must agree to the participation of each individual even before a solution has been

developed. In many situations (particularly involving humans), it is difficult to determine before a solution has actually been developed whether an individual is capable of contributing. Therefore a more pragmatic approach may be to weaken this condition and allow agents to participate in the subsequent solution development phase on the basis that they alone believe they can contribute. This means at the outset of group formation, agents which have no possible means of contributing can be ruled out and then at a later stage when the actual solution is developed superfluous agents can be removed.

3.4 And Finally: Joint Responsibility

We are now in a position of being able to describe the mental state which a group of agents must adopt if they are to jointly solve a common problem:

$$\textbf{\textit{JOINT-RESPONSIBILITY}}\ (\langle\{\alpha_1,..,\alpha_n\}, \sigma\rangle) \Leftrightarrow$$
$$\textbf{\textit{MB}}\ (\{\alpha_1,..,\alpha_n\}, \textbf{\textit{JPG}}(\langle\{\alpha_1,..,\alpha_n\}, \sigma\rangle)) \wedge$$
$$\textbf{\textit{MB}}\ (\{\alpha_1,..,\alpha_n\}, \textbf{\textit{SOLUTION-COMMITMENT}}\ (\langle\{\alpha_1,..,\alpha_n\}, \sigma\rangle)) \wedge$$
$$\textbf{\textit{MB}}\ (\{\alpha_1,..,\alpha_n\}, (\forall\ \alpha_i \in \{\alpha_1,..\alpha_n\}\ \textbf{\textit{MAY-CONTRIBUTE}}(\alpha_i, \langle\{\alpha_1,..,\alpha_n\}, \sigma\rangle))$$

4. CONCLUSIONS

The work presented in this paper is a synthesis and extension of previous work in the fields of multi-agent planning and joint intentions and provides a foundation upon which robust and sophisticated collaborative problem solving can be based. Joint responsibility offers, for the first time, a model which defines the pre-conditions which need to be satisfied before joint action can commence as well as a prescription of how agents should behave when engaged in collaborative problem solving. The theory as described assumes that the group is formed dynamically, but covers already existing organizational structures as a special case. A code of conduct for group problem solving is particularly important in situations in which agents' plan and goal states alter dynamically (eg in complex real world environments [13]) or when engaged in prolonged social activity.

Responsibility is also capable of providing a degree of trust upon which all social interactions must be based [2]: team members can carry out problem solving activity safe in the knowledge that others will be doing their bit, and if they are not then they will at least be endeavouring to inform other team members of this fact. Such assurances mean that as far as can be avoided, agents will not be wasting valuable computational effort pursuing activities which will ultimately serve no purpose, because group actions often only make sense when performed within the context of the group's activities.

5. REFERENCES

[1] J.F.Allen (1984) "*Toward a General Theory of Time and Action*" Artificial Intelligence, 23, pp 123-154

[2] R.Axelrod (1984), "*The Evolution of Cooperation*", Basic Books Inc. New York

[3] M.E.Bratman, (1990), "*What is Intention?*" Intentions in Communication, (eds P.R.Cohen, J.Morgan & M.E.Pollack), pp 15-33, MIT Press

[4] P.R.Cohen & H.J.Levesque, (1990), "*Intention is Choice with Commitment*" Artificial

Intelligence, 42, pp 213-261.

[5] E.H.Durfee & V.R.Lesser, (1987), "*Using Partial Global Plans to Coordinate Distributed Problem Solvers*", Proc IJCAI 1987, pp 875-883.

[6] J.R.Galliers, (1989) "*The Positive Role of Conflict in Cooperative Multi-Agent Systems*", Proc MAAMAW 1989, Cambridge, UK.

[7] B.J.Grosz & C.L.Sidner, (1990), "*Plans for Discourse*", Intentions in Communication, (eds P.R.Cohen, J.Morgan & M.E.Pollack), pp 417-444, MIT Press

[8] D.Harel (1979) "*First Order Dynamic Logic*", New York, Springer Verlag.

[9] J.Hendler, A.Tate & M.Drummond (1990) "*AI Planning: Systems and Techniques*", AI Magazine, pp 61-77

[10] J.R.Hobbs, (1990), "*Artificial Intelligence and Collective Intentionality*", Intentions in Communication, (eds P.R.Cohen, J.Morgan & M.E.Pollack), pp 445-460, MIT Press.

[11] H.J.Levesque, P.R.Cohen & J.Nunes, (1990), "*On Acting Together*", Proc AAAI, 94-99.

[12] N.R.Jennings, (1991) "*ARCHON: An Architecture for Cooperating Systems*" in Artificial Intelligence and Simulation of Behaviour Quarterly, Special Issue on Distributed AI.

[13] N.Jennings (1991) "*Cooperation in Industrial Systems*" ESPRIT Conference, 253-263.

[14] K.E.Lochbaum, B.J.Grosz & C.L.Sidner, (1990), "*Models of Plans to Support Communication*", Proc AAAI 90, pp 485-490.

[15] G.H.Mead, (1934), "*Mind, Self and Society*", Univ. of Chicago Press, Chicago.

[16] A.S.Rao & M.P.Georgeff, (1991), "*Social Plans: A Preliminary Report*", in this volume.

[17] J.R.Searle, (1990), "*Collective Intentions and Actions*", in Intentions in Communication, (eds P.R.Cohen, J.Morgan & M.E.Pollack), pp 401-416, MIT Press

[18] M.P.Singh, (1990), "*Group Ability and Structure*" Proc. MAAMAW 1990, France.

[19] R.G.Smith & R.Davis, (1981), "*Frameworks for Cooperation in Distributed Problem Solving*", IEEE SMC, 11, 1, pp 61-70

[20] R.Tuomela & K.Miller, (1988), "*We-Intentions*", Philosophical Studies 53, 367-389.

[21] C.Roda, N.R.Jennings & E.Mamdani, (1990), "*ARCHON: A Cooperation Framework for Industrial Process Control*", in Cooperating Knowledge Based Systems, (ed S.M.Deen) pp 95-112, Springer Verlag.

[22] J.Rosenschein, (1982), "*Synchronization of Multi-Agent Plans*", AAAI, pp 115-119.

[23] J.Rosenschein & M.Genesereth, (1985), "*Deals Among Rational Agents*", IJCAI, 91-99.

[24] E.Werner, (1989), "*Cooperating Agents: A Unified Theory of Communication & Social Structure*", in Distributed Artificial Intelligence Vol II, (eds Gasser & Huhns), pp 3-36.