Commitments and Conventions: The Foundation of Coordination in Multi-Agent Systems

NICK R. JENNINGS

Department of Electronic Engineering, Queen Mary and Westfield College, University of London, Mile End Road, London E1 4NS, UK.

[email: N.R.Jennings@qmw.ac.uk]

Abstract

Distributed Artificial Intelligence systems, in which multiple agents interact to improve their individual performance and to enhance the system's overall utility, are becoming an increasingly pervasive means of conceptualising a diverse range of applications. As the discipline matures, researchers are beginning to strive for the underlying theories and principles which guide the central processes of coordination and cooperation. Here agent communities are modelled using a distributed goal search formalism and it is argued that *commitments* (pledges to undertake a specified course of action) and *conventions* (means of monitoring commitments in changing circumstances) are the foundation of coordination in multi-agent systems. An analysis of existing coordination models which use concepts akin to commitments and conventions is undertaken before a new unifying framework is presented. Finally a number of prominent coordination techniques which do not *explicitly* involve commitments or conventions are reformulated in these terms to demonstrate their compliance with the central hypothesis of this paper.

1 Introduction

As more challenging applications are automated, so the size and complexity of software systems becomes greater. However this process cannot continue indefinitely; already researchers from a variety of disciplines have noted that present generation intelligent systems are nearing the

boundaries of current software engineering approaches and that a fundamental shift of paradigm is required (Cox, 1990; Lenat and Feigenbaum, 1991; McDermott, 1990; Stefik, 1986). Distributed Artificial Intelligence (DAI) is one of these new approaches. It aims to construct systems composed of multiple problem solving entities which interact with one another to enhance their performance. With this "divide and conquer" approach the scope of each component is limited, meaning the complexity of the computation is lower, thus enabling the processing elements to be simpler and more reliable. This increased demand on software systems has also coincided with important advances in hardware technology for processor fabrication and for inter-processor communication - meaning it is now economically feasible and technically viable to connect together large numbers of powerful, yet inexpensive, processing units that execute asynchronously.

Decentralised and cooperative problem solving systems have been advocated as a means of: increasing the level of information integration across organisations (Pan and Tenenbaum, 1991; Papazoglou *et al.*, 1992; Shina, 1991); overcoming the limitations on intelligence present in any finite artificial system (March and Simon, 1958; Minsky 1985; Simon, 1957); developing sophisticated applications (Jennings and Wittig, 1992; Neches *et al.*, 1991) and providing a more natural representation of distributed problems[1] (eg sensor networks (Lesser and Corkill, 1983), air traffic control (Cammarata *et al.*, 1983), information retrieval (Huhns *et al.*, 1988) and electricity networks (Jennings *et al.*, 1992)). Other potential advantages include: reusability of problem solving components by incorporating the same system into several cooperating communities, an increased set or scope of achievable tasks by sharing resources, improved system robustness by undertaking duplicate tasks using different methods, enhanced problem solving due to the combination of multiple problem solving paradigms and sources of information and problem solving speed up due to parallel execution (Bond and Gasser, 1988; Durfee *et al.*, 1987; Gasser and Huhns, 1989; Huhns, 1988).

[1.] It has even been suggested that all real problems are distributed (Hayes-Roth, 1980)

In DAI systems, agents are grouped together to form communities which cooperate to achieve the goals of the individuals and of the system as a whole. This review concentrates on agents which possess a range of identifiable problem solving capabilities, have their own aims and objectives, are relatively autonomous in deciding what actions to perform and can reason about the process of coordination. These characteristics debar some well known networks of cooperating entities including ACTORS (Agha, 1986), BEINGS (Lenat, 1975) and neural networks (McClelland and Rumelhart, 1986). In all of these systems, cooperative behaviour stems from predefined interactions between tightly coupled, simple processing elements. Each individual has little knowledge of the system's overall objective or of general strategies for communication and coordination. Therefore the agents cannot perform meaningful problem solving in their own right nor can they operate outside of the specific cooperation protocols specified in advance by the system designer.

In the majority of multi-agent systems, community members have problem solving expertise which is related, but distinct, and which frequently has to be coordinated when solving problems. Such interactions are needed because of the dependencies between agents' actions, the necessity of meeting global constraints and because no one individual has sufficient competence, resources or information to solve the entire problem.

Interdependence occurs when goals undertaken by individual agents are related - either because local decisions made by one agent have an impact on the decisions of other community members (eg when building a house, decisions about the size and location of rooms impacts upon the wiring and plumbing) or because of the possibility of harmful interactions amongst agents (eg two mobile robots may attempt to pass through a narrow exit simultaneously, resulting in a collision, damage to the robots and blockage of the exit).

Global constraints exist when the solution being developed by a group of agents must satisfy certain conditions if it is to be deemed successful. For instance a house building team may have a budget of £250,000, a distributed monitoring system may have to react to critical events within 30

seconds and a distributed air traffic control system may have to control the planes with a fixed communication bandwidth. If individual agents acted in isolation and merely tried to optimise their local performance, then such overarching constraints are unlikely to be satisfied. Only through coordinated action will acceptable solutions be developed.

Finally many problems cannot be solved by individuals working in isolation because they do not possess the necessary expertise, resources or information. Relevant examples include the tasks of lifting a heavy object, driving in a convoy and playing a symphony. It may be impractical or undesirable to permanently synthesize the necessary components into a single entity because of historical, political, physical or social constraints, therefore temporary alliances through cooperative problem solving may be the only way to proceed. Differing expertise may need to be combined to produce a result outside of the scope of any of the individual constituents (eg in medical diagnosis, knowledge about heart disease, blood disorders and respiratory problems may need to be combined to diagnose a patient's illness). Different agents may have different resources (eg processing power, memory and communications) which all need to be harnessed to solve a complex problem. Different agents may have different information or viewpoints of a problem (eg in concurrent engineering systems, the same product may be viewed from a design, manufacturing and marketing perspective).

Even when individuals can work independently, meaning coordination is not essential, information discovered by one agent can be of sufficient use to another that the two agents can solve the problem more than twice as fast. For example when searching for a lost object in a large area it is often better, though not essential, to do so as a team. Analysis of this "combinatorial implosion" phenomena (Hewitt and Kornfield, 1980) has resulted in the postulation that cooperative search, when sufficiently large, can display universal characteristics which are independent of the nature of either the individual processes or the particular domain being tackled (Clearwater *et al.*, 1991).

This paper does not aim to provide comprehensive coverage of the entire field of DAI - such

reviews can be found in (Bond and Gasser, 1988; Chaib-Draa *et al.*, 1992; Decker, 1987; Durfee *et al.*, 1989; Gasser, 1991/92a; Hern, 1988). Rather the objective is to carry out an in-depth analysis on work related to coordinating the problem solving of multiple agents which is one of the central problems of DAI research. At present there are a diverse range of techniques which can and do facilitate coordination in DAI systems. However these mechanisms vary considerably in their time horizon, the level of predictive information they provide, the computational overhead they require and the assumptions they make about an agent's architecture and mental state.

To develop better and more integrated models of coordination, and hence improve the efficiency and utility of DAI systems, it is necessary to obtain a deeper understanding of the fundamental concepts which underpin agent interactions. Here a distributed goal search formalism is used to characterise DAI systems and a unifying coordination model is presented which has the notions of *commitment* and *convention* at its core (section three). Commitments are viewed as pledges to undertake a specified course of action, while conventions provide a means of monitoring commitments in changing circumstances. The former provide a degree of predictability so that agents can take the (future) activities of others into consideration when dealing with inter-agent dependencies, global constraints or resource utilization conflicts. The latter provide the flexibility which cooperating agents need if they are to cope with being situated in dynamic environments. To operate effectively when the external world and their own beliefs are constantly changing, agents must possess a mechanism for evaluating whether existing commitments are still valid. Conventions provide this mechanism: defining the conditions under which commitments should be re-assessed and specifying the associated actions which should be undertaken in such situations.

This new model of coordination is founded upon the "*Centrality of Commitments and Conventions Hypothesis*" which states that: *all coordination mechanisms can ultimately be reduced to (joint) commitments and their associated (social) conventions*. To provide a context for the new framework a number of extant coordination models, which use concepts akin to commitments and conventions, are discussed (section two). This review identifies the important

intuitions which need to be captured and highlights the inconsistent and often informal way in which the key notions are presently used. Finally section four investigates three prominent models of coordination (organisational structuring, meta-level information exchange and multi-agent planning) which do not make *explicit* use of commitments or conventions, and shows how they can all be reformulated in these terms - thus providing further evidence for the main claim of this paper.

2 An Analysis of Commitments and Conventions in Extant Coordination Models

Participation in any social situation should be both simultaneously constraining, in that agents must make a contribution to it, and yet enriching, in that participation provides resources and opportunities which would otherwise be unavailable (Gerson, 1976). Coordination, the process by which an agent reasons about its local actions and the (anticipated) actions of others to try and ensure the community acts in a coherent manner, is the key to achieving this objective. Without coordination the benefits of decentralised problem solving vanish and the community may quickly degenerate into a collection of chaotic, incohesive individuals. Coordination aims to ensures that all necessary portions of the overall problem are included in the activities of at least one agent, that agents interact in a manner which permits their activities to be developed and integrated into an overall solution, that team members act in a purposeful and consistent manner and that all of these objectives are achievable within the available computational and resource limitations (Lesser and Corkill, 1987). Specific examples include supplying timely information to needy agents, ensuring actions are synchronised and avoiding redundant problem solving.

When viewing agents from a purely behaviouristic (external) perspective, it is, in general, impossible to determine whether they have coordinated their actions. Firstly actions may be incoherent even if the agents tried to coordinate their behaviour. This may occur, for example, because their models of each other or of the environment are incorrect. For example, $robot_1$ may see $robot_2$ heading for $exit_2$ and, based on this observation and the subsequent deduction that it will use this exit, decide to use $exit_1$. However if $robot_2$ is heading towards $exit_2$ to pick up a particular item and actually intends to use $exit_1$ then there may be incoherent behaviour (both agents

attempting to use the same exit) although there was coordination. Secondly even if there is coherent action, it may not be as a consequence of coordination. For example imagine a group of people are sitting in a park (Searle, 1990). 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 uncoordinated behaviour because 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 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 it is coordinated behaviour because they each have the aim of meeting at the central point as a consequence of the overall aim of executing the dance. For these two reasons, coordination is best studied by examining the mental state of the individual agents. The exact make up of this mental state is still the subject of much debate, however there is an emerging consensus on the fact that it contains beliefs, desires, goals and commitments (intentions).

If all the agents could have complete knowledge of the goals, actions and interactions of their fellow community members and could also have infinite processing power, it would be possible to know exactly what each agent was doing at present and what it is intending to do in the future. In such instances, it would be possible to avoid conflicting and redundant efforts and systems could be perfectly coordinated (Malone, 1987). However such complete knowledge is infeasible, in any community of reasonable complexity, because bandwidth limitations make it impossible for agents to be constantly informed of all developments. Even in modestly sized communities, a complete analysis to determine the detailed activities of each agent is impractical - the computation and communication costs of determining the optimal set and allocation of activities far outweighs the improvement in problem solving performance (Corkill and Lesser, 1986).

As all community members cannot have a complete and accurate perspective of the overall system, the next easiest way of ensuring coherent behaviour is to have one agent with a wider

picture. This global controller could then direct the activities of the others, assign agents to tasks and focus problem solving to ensure coherent behaviour. However such an approach is often impractical in realistic applications because even keeping one agent informed of all the actions in the community would swamp the available bandwidth. Also the controller would become a severe communication bottleneck and would render the remaining components unusable if it failed.

To produce systems without bottlenecks and which exhibit graceful degradation of performance, most DAI research has concentrated on developing communities in which both control and data are distributed. Distributed control means that individuals have a degree of autonomy in generating new actions and in deciding which tasks to do next. When designing such systems it is important to ensure that agents spend the bulk of their time engaged on solving the domain level problems for which they were built, rather than in communication and coordination activities. To this end, the community should be decomposed into the most modular units possible. However the designer should ensure that these units are of sufficient granularity to warrant the overhead inherent in goal distribution - distributing small tasks can prove more expensive than performing them in one place (Durfee *et al.*, 1987; Wesson *et al.*, 1981).

The disadvantage of distributing control and data is that knowledge of the system's overall state is dispersed throughout the community and each individual has only a partial and imprecise perspective. Thus there is an increased degree of uncertainty about each agent's actions, meaning that it more difficult to attain coherent global behaviour - for example agents may spread misleading and distracting information, multiple agents may compete for unshareable resources simultaneously, agents may unwittingly undo the results of each others activities and the same actions may be carried out redundantly. Also the dynamics of such systems can become extremely complex, giving rise to nonlinear oscillations and chaos (Huberman and Hogg, 1988). In such cases the coordination process becomes correspondingly more difficult as well as more important[2].

To ensure agents can reason about how their actions will contribute to the collective problem solving effort and how they can benefit from interactions with others, a number of models of

coordination have been developed. This section concentrates on those models which explicitly involve concepts similar to commitments and conventions - initially reviewing models which define individual behaviour (section 2.1), before moving onto models for social behaviour (section 2.2).

2.1 Models of Individual Behaviour

There have been many attempts to define the behaviour of rational problem solvers through the notion of *intentions* (Becker, 1960; Bratman, 1984; Dennett, 1987; Searle, 1983). Like many other folk psychology concepts, the term "intention" has been given various interpretations, although most of them have the concept of commitment at their core. The notion of commitment is given such prominence because it is central to three aspects of an agent's practical reasoning process (Bratman, 1984). Firstly as agents are resource bounded, they cannot continually weigh their competing desires and their associated beliefs in deciding what to do next. At some point, the agent must just settle on one state of affairs for which to aim, thus creating a *commitment* to obtain that objective. Secondly commitments are needed to plan and coordinate future actions. Once a future action has been decided upon, the agent must make subsequent decisions within the context that it will perform the said action. In a social environment, agents use knowledge of their acquaintances' commitments to determine what actions they will perform so that their own actions can be organised to achieve the best results for themselves and for the community. Finally, commitments pose problems for means-end analysis; they provide a high-level goal for which the agent must find a suitable course of action.

However as well as embodying commitment, many of the definitions of intentions also intertwine notions about tracking commitments (conventions) so that the two concepts become

---

[2.] Similar experiences have also been noted in organisational science: the greater the task uncertainty, the greater the amount of information which must be processed among decision makers during task execution in order to achieve a given level of performance (Galbraith, 1973).

virtually indistinguishable. To clarify the situation, several prominent models of individual intentions are reviewed and placed within the commitment plus convention framework.

The most comprehensive attempt to formalise individual intentions is due to Cohen and Levesque (1990). Their model has two levels of detail; a fundamental one which provides the primitives for a theory of action (eg definitions of beliefs, goals and action) and a second layer which builds upon these concepts to develop a theory of rational action. At the second layer they capture the notion of commitment by defining a *persistent goal*. Agent A has a persistent goal to achieve objective G, relative to its motivation M, if and only if the following conditions prevail: (i) A believes that G is currently false; (ii) A wants G to be eventually true; (iii) this state of affairs will continue until A comes to believe either that G is true or that it will never be true or that M is false[3]. An intention is then defined as a commitment to act in a certain mental state: agent A *intends* to do action G if it has the persistent goal to have done that action and, moreover, to have done it believing throughout that it was doing it.

Persistent goals embody ideas related to both commitments and conventions, which has led to claims that their definition is circular (Singh, 1992). This problem could be avoided if the two concepts were separated out - commitments defined as a primitive notion and then associated conventions specified to operate on them (see figure 1 for an example). The close interweaving of commitments and notions of rationality restricts the generality of the system. For example, it is not possible to model an agent which fanatically follows a goal until it is satisfied. This is because their "realism" constraint on the semantics of the underlying GOAL operator requires that if an agent comes to believe that it will never achieve its objective, then it must drop its persistent goal (i.e. under no circumstances can it remain committed). Adopting the advocated demarcation

[3]. Other authors have suggested different situations in which commitments may be dropped - these include allowing agents to keep goals so long as they believe the objective is still viable and allowing agents to remain committed so long as they still desire the objective (Rao and Georgeff, 1991), or allowing commitments to be dropped if more highly rated, but conflicting, alternatives become available (Galliers, 1988).

would also mean that the notion of commitment remains consistent in all situations; different types of social system can then be represented simply by changing the relevant convention.

```
CONVENTION: Cohen-Levesque-Model

REASONS FOR RE-ASSESSING COMMITMENT:
   Commitment by agent A to goal G relative to motivation M
           • A BELIEVES G IS TRUE
           • A NO LONGER BELIEVES G WILL BECOME TRUE
           • A NO LONGER HOLDS MOTIVATION M

ACTIONS:

   R1:   IF COMMITMENT SATISFIED OR
            COMMITMENT UNATTAINABLE OR
            MOTIVATION FOR COMMITMENT NO LONGER PRESENT
         THEN CONSIDER DROPPING COMMITMENT
```

Figure 1: Convention for Cohen and Levesque's Model

Pollack's (1990) model of problem solving has two components, one related to an agent's beliefs and another to its intentions. In this work, intentions are given no finer structure nor formal semantics - they are merely defined in an intuitive manner without reference to notions such as commitment, stability or consistency. For agent A to have a plan to (intend to) do G that consists of doing actions P, A must have the following beliefs: (i) A must believe that executing the actions in P, in their specified temporal order, will entail performance of G; (ii) A must believe that each p_i in P plays a role in the plan. Thus A believes that by doing p_i *it will do* G or some other action Q that plays a role in its plan, or that doing p_i will *enable doing* G or some other Q that plays a role in the plan. The former corresponds to generator relationships and the latter to enablement. If action a *generates* action b then the agent only need do a and b will automatically be done. However when a *enables* b, the agent needs to do something more than a to guarantee that b will be done. For example, knowing the phone number of a pizza store enables a pizza to be ordered, if there is access to a phone, but it does not generate the order.

Pollack claims these beliefs are necessary but not sufficient conditions to guarantee that doing

P is A's plan to do G. For sufficiency, it is also necessary for A to have a certain set of "intentions" with respect to P. In particular A must intend: (iii) to execute each p_i in P in the specified temporal order; (iv) to execute P as a way of doing G (this circumvents the problem of accidently carrying out an action which succeeds in bringing about the desired outcome); (v) that each p_i in P plays a role in the plan (A must intend by doing p_i to *do* G or some other Q that plays a role in its plan or by doing p_i to *enable* doing G or some other Q that plays a role).

There is clearly a close link between the second and fifth conditions of this formulation. The distinction between merely knowing of a sequence of steps which will achieve a particular objective and actually intending to follow this sequence ensures that having the beliefs described in condition (ii) does not imply having the intention described in (v). However it is unclear as to whether having the intention in condition (v) necessarily means holding the beliefs in condition (ii). There is literature to support both views: (a) plans normally support expectations of their successful execution (Audi, 1973); (b) I may intend to make ten legible copies of what I am writing by pressing hard on carbon paper, without believing with any confidence that I may succeed (Davidson, 1980). Adopting the former view means that (v) directly entails (ii), and also that (iv) entails (i); adopting the latter requires both aspects to be explicitly present. Pollack adopts the latter view because it is useful when an inferring agent deems an actor's plan invalid to determine whether this is through belief in the plan action *per se* or whether it is due to incompatible beliefs.

Although this work represents an important contribution to the field of plan inference, in that it stresses the central role of mental attitudes alongside plan structure, from the view of controlling agent activity it contains many flaws. Most importantly, the notion of intention is stated only informally, and the interpretation which most readily springs to mind is that of commitment. There is no notion of convention - thus the model does not adequately explain how an agent should behave if things go wrong. For example, if an agent no longer believes it is capable of executing an action, what should it do? Should it give up? Should it replan? This

formalism, simply states that the agent no longer has P as a plan.

Werner (1989) outlines a general theoretical framework for designing agents with a communicative and social competence. An important aspect of his agent model is the idea of an "intentional state" which represents the set of strategies an agent can use to guide its actions. This intentional state can be abstractly defined by the use of social roles. For example in a master-slave interaction, the role of the master contains an expectation that it will tell the slave what to do and the role of the slave carries the expectation that it will follow the master's instructions. Although roles are clearly a form of commitment, in that they represent a pledge to act in a certain manner, it is unclear exactly what agents are committing themselves to since there is no representation of either a goal or a common plan. Hence it is difficult to determine what it means for an agent to employ a particular strategy in both theoretical and computational terms. As another example, the contract-net cooperation protocol (Smith and Davis, 1981), in which agents bid to undertake tasks advertised by a manager node, has two roles: $\text{rol}_{\text{manager}}$ and $\text{rol}_{\text{contractor}}$. Again Werner only provides a textual description of what the roles are and what it means to adopt a particular role, meaning that it is difficult to assess what level of commitment is implied by undertaking a particular role. Although roles are designed to be used in social environments they only specify the mental states of individuals, there is no notion of group. Finally no mechanisms are provided for reassessing commitments, meaning that once a role is adopted it must be adhered to indefinitely.

The first two models concentrate on defining commitments and conventions which are applicable in asocial situations, while the third model defines a purely individualistic perspective on social actions. This work is relevant because in any multi-agent system there will be some goals which are worked on by individuals and which will be unrelated to the activities of others. In such cases, asocial commitments and conventions are sufficient for describing an agent's behaviour. However to express the full richness of interactions which are possible in a social context, formulations specifically conceived for collaborative problem solving are required

(Gilbert, 1989; Power, 1984).

2.2 Models of Social Behaviour

This subsection investigates several coordination models which explicitly deal with concepts such as joint goals, joint actions, joint commitments or social conventions. The models are divided into three main categories: (i) formal models which describe social actions from the viewpoint of the participating individuals; (ii) formal models which describe social actions using descriptions of teams and joint goals as primitive concepts; (iii) computational models of social behaviour.

2.2.1 Formal Models, Individualistic Perspective

Grosz and Sidner (1990) propose a special operator, called a SharedPlan, for describing collaborative problem solving between a group of agents a_1,.... a_n which are attempting to achieve a particular objective G. SharedPlans require the following to be mutually believed[4] for each subgoal of G: (i) one team member (a_i ε a_1,....a_n) is capable of executing the action; (ii) a_i "intends" to achieve the subgoal; (iii) a_i intends to achieve G "BY" performing the subgoal. Agents also need to have mutual belief about the generator relationships (see section 2.1) between sub-goals and how they lead to achievement of the parent goal. An example is that the sub-goals of lifting at opposite ends of a heavy object will result in (generate) achievement of the overarching goal of lifting that object if they are carried out simultaneously.

In this formalism the notion of intention is not rigourously defined, intuitively it is used to represent the concept of commitment. Even ignoring this shortcoming, a number of conceptual problems still remain. Firstly the important notion of BY, which links a subgoal to its parent, is given no formal semantics and can result in counter intuitive observations. In the description of a collaborative lift the following clause appears: INTEND(a_1, BY(lift(end_1), lift(heavy-object))),

4. Mutual belief is the infinite conjunction of beliefs about other agents' beliefs about other agents' beliefs and so on to any depth about some proposition (Halpern, 1986).

there is a similar clause for a_2 lifting at end_2. This statement is clearly nonsense, the heavy object will not be moved by a_1 lifting at end_1, rather it will only happen if a_2 also lifts at end_2 simultaneously. This problem arises because of the lack of the overarching concept of a joint goal, a possibility explicitly ruled out by the insistence that agents can only intend their own actions. Secondly a myriad of different generator relationships are required for each and every possible interrelationship between goals and sub-goals - one for simultaneous actions, one for conjoined actions, one for sequences of actions, and so on.

In a refinement of this work, BY is replaced by a "CONTRIBUTES" relation and generator relationships are replaced by mutually believed sequences of action with a known outcome (Lochbaum *et al.*, 1990). Action sequences can encode various goal interrelationships and remove the need to produce different generator relationships for each type of constraint. CONTRIBUTES allows a more natural expression of goal-subgoal relationships. Returning to the collaborative lift example, a_1 believes that: INTEND(a_1, lift(end_1)) and that CONTRIBUTES(lift(end_1), a_1, lift(heavy-object)). Although these modifications are an improvement, the formalism does not allow joint goals to be represented, nor is there any indication of when commitments should be reassessed and how to act towards others if the decision is to renege.

Tuomela and Miller (1988) propose we-intentions (eg "we shall do G") as a means of describing collaborative situations. They believe that in order to study social action, it is necessary to have a clear idea about the internalisation of the notion of "group" in its members. We-Intentions are the basis of the sociality inherent in acting together; thus if an agent intentionally performs a helpful act, but does not share the relevant group intention expressing the overall common goal, then it cannot be considered as part of the joint action. Agent a_i is a member of a group which we-intends to do G if: (i) a_i intends to do its part of G; (ii) a_i believes that the joint action opportunities for G are true, especially that at least a sufficient number of the full-fledged and adequately informed members of the group, as required for the performance of G, will do their part; (iii) a_i believes there is mutual belief amongst group members to the effect that the

preconditions of success mentioned above hold true.

As with previous formalisms, intends is informally used to represent the notion of commitment. The second component of the we-intentions definition highlights the shortcoming of Lochbaum *et al.*'s "CONTRIBUTES" relation. Unlike the SharedPlan relation, it expresses the strong interdependence of individual intentions when agents are working together and it also highlights the crucial role which commitments play in social actions. Actions only contribute to the overall objective when they are considered in conjunction with those of others, there can be no absolute contribution to a group action if strong inter-goal dependencies exist. Commitments ensure there is sufficient trust in the belief that the other agents will do their part for an individual to fulfill its part of G.

Like the SharedPlan work, this analysis attempts to reduce collective behaviour to individual commitments plus beliefs. There is no attempt to explicitly represent the total social action G, only the subparts performed by the agent in question appear in the definition. Such formulations are subject to counter examples of the following form: consider a musician in an orchestra who intends to perform his part of the overall action properly, but nevertheless intends to do something which will make the visiting conductor look ridiculous and will spoil the orchestra's overall performance. Such a musician cannot be said to we-intend to play the symphony even though it intends to perform an act which brings the shared objective closer. Rather than make use of the notion of a joint intention, in which this problem could not arise, Tuomela and Miller place a very strong interpretation on the "intends to do its part" aspect of the formulation. They require that an not only agent accepts "I shall do my part of G" as being true of itself, but also the stronger statement that "we shall do G". The latter part of this definition is not formalised.

Although Tuomela and Miller acknowledge that commitments can be broken and that this can affect the whole group; they simply state that when things go wrong with one agent's activities, the other group members will help exert pressure and do whatever they think is necessary for the collective to succeed in achieving its objective. They have no convention which defines "go

wrong" nor any conceptualisation of "do whatever they think is necessary for the collective to succeed."

Cohen and Levesque (1991a) formulate joint commitment through the definition of joint persistent goals which, in turn, are based upon the concept of *achievement goals*. Achievement goals define the state of individuals participating in a team which is working towards a common objective with a specified motivation. Agent a_i has a *weak achievement goal*, relative to its motivation M, to bring about G if either of the following are true: (i) a_i does not yet believe that G has been achieved and has G being eventually true as a goal (i.e. a_i has a *normal achievement goal* to bring about G); (ii) a_i believes that G is true, will never be true or is irrelevant (M is false), but has a goal of making the status of G mutually believed by all team members. A team of agents has a joint persistent goal, relative to M, to achieve G if and only if: they mutually believe that G is currently false; they mutually believe that they all want G to be eventually true and until they come to mutually believe either that G is true, that G will never be true or that M is false, they will continue to mutually believe that they each have G as a weak achievement goal relative to M[5].

If a team is jointly committed to achieving G, they mutually believe that they each have G as a normal achievement goal initially. However as time passes, team members cannot rely on the fact that they all still have G as a normal achievement goal; they can only assume that they have it as a weak achievement goal. The reason for this weaker statement is that one team member may have discovered that the goal is finished (impossible or irrelevant) and may be in the process of making this fact known to its associates. If at some point it is no longer mutually believed that everybody has the normal achievement goal, then there is no longer a joint persistent goal as not all the agents wish G to be true. Thus the team is no longer jointly committed to G. However a weak achievement goal persists and ensures that all team members are informed of the lack of

5. In their account of confirmations in task-oriented dialogues, a number of slight variations in the definitions of both individual and joint commitments are presented (Cohen and Levesque, 1991b). Also included in this work are a variety of conventions for monitoring the execution of joint intentions.

commitment by the doubting individual. This means agents can rely upon the commitments of others; firstly to the overall objective and then, if necessary, to the mutual belief of the status of the objective (Levesque *et al.*, 1991). Individuals can therefore undertake activities in the knowledge that others are working towards the same overall objective and that if something goes awry then they will be informed. A joint intention for a team of agents to achieve G, relative to their motivation M, occurs when all the members have a joint persistent goal relative to M of their having done G and, moreover, having done it mutually believing throughout that they were doing it.

Like their work on individual intentions, the concept of a joint persistent goal contains both the notion of commitment and also a convention which explains how to monitor it. The model of commitment hardwires a set of conditions under which commitments can be dropped and specifies a definitive code of conduct for how to behave towards others in the team when commitments are reneged upon. A further disadvantage of this model is that there is no explicit representation of collectives - actions and goals are defined solely in terms of individuals.

2.2.2 Formal Models, Societal Perspective

Many of the problems with the models discussed in the previous sub-section occur because they attempt to define cooperative behaviour in terms of individual goals. An alternative approach adopted by several researchers is to acknowledge that joint goals are a primitive concept in their own right. That is, joint goals cannot be analysed *solely* in terms of individual goals, even if the individual goals are supplemented with beliefs about the related goals of other agents[6]. The reason for this belief is that the former imply the notion of cooperation whereas the latter do not. Even if two agents possess the same individual goal, and they are both mutually aware of this fact,

[6.] This position is consistent with the symbolic interactionist school of sociological thought which adheres to the view that joint action is the fundamental unit of society (Mead, 1934). A more traditional AI statement of this problem is that an agent can have as goals in its plan, logical formulae whose predicates describe actions that collectives engage in and whose agent argument is such a collective (Hobbs, 1990).

this does not entail the presence of a desire to cooperate.

Rao *et al.* (1992) adopt this approach, augmenting the notions of individual intentions, beliefs and desires with structures describing joint goals and joint intentions. Their basic unit of activity is the plan expression, a pair consisting of a plan type (an abstract structure that, when executed by an agent, results in the occurrence of an action in the real world) and an agent. Complex actions involving goal interdependencies are obtained by combining plan expressions with operators from dynamic logic (eg sequence, parallelism and non-deterministic choice (Harel, 1984)). In addition to describing the activities of individuals it is also possible to seamlessly define social actions, as the "agent" in a plan expression can correspond to a team. This enables joint goals and joint commitments to be represented in a simple and elegant manner. A joint goal G is defined as meaning that all members of the group have the same goal G and that they all mutually believe that G is held as a joint goal. Joint intentions are defined in the same way. Following from these definitions, they prove a theorem which states that: if a group of agents jointly intends a particular action, they all have it as a joint goal and also mutually believe it. Kinny *et al*. (1992) have adapted and extended this formalism to allow joint plans to be expressed at a team level and have also shown how these plans can be used to guide the activities of the individual agents.

This formalism adequately represents the notion of commitment to joint activity and the theorem relating joint intentions and joint goals captures, in a formal framework, the inherent notion of cooperation present in collaborative actions. By defining joint intentions independently of considerations about monitoring commitments, the formalism offers an opportunity to represent social systems which have different types of conventions. However despite recognising that actions may fail, and hence commitments may be reneged upon, the formulation does not embody any form of convention which describes how to react in such circumstances. Also by opting for dynamic logic to combine plan operators, the expressiveness of the planning language is limited. A general purpose goal interdependence operator would remove this shortcoming.

The model of Joint Responsibility (Jennings, 1992) also describes joint action in terms of teams of agents. This model extends Cohen and Levesque's work on joint persistent goals (see 2.2.1) and stipulates that agents should make joint commitments to agreed sequences of actions as well as to the shared objective itself. Commitment to the common solution provides a context for the performance of actions in much the same way as the shared aim guides the objectives of the individuals. The convention for joint persistent goals defines how to monitor commitments to the common objective and an additional convention specifies that the common plan should be re-examined if any of the following conditions arise: the agreed plan will not achieve the desired results, the agreed plan cannot be executed or the agreed plan has not been executed properly. The model has been implemented in a general purpose cooperation framework and applied to the real-world problem of electricity transportation management where it led to high degrees of coordination even in the most unpredictable and dynamic situations (Jennings and Mamdani, 1992).

2.2.3 Computational Models

At present there is a relatively large gap between the theoretical models of commitments and conventions which were described in the previous two sub-sections and implemented DAI systems. This chasm exists partly because of the differing motivations of model and system builders, but also reflects some major theoretical shortcomings. For example most of the theoretical models embody the notion of mutual belief, however it has been shown that this is unattainable in systems in which communication is not guaranteed or when there is some uncertainty in message delivery time (Halpern and Moses, 1984).

Bratman *et al.* (1988) have devised a Belief-Desire-Intention architecture in which commitments play a central role in guiding an agent's actions and future planning and in which an "override mechanism" provides a computational realisation of conventions. This proposal is predominantly for an individual agent situated in an asocial context, but it provides a useful functional architecture which can be augmented for use in a cooperative context. Burmeister and

Sundermeyer (1992) specify and implement an architecture based on individual intentions which takes into account the fact that agents are situated in multi-agent environments. Their representation of intention, although not based on a particular theoretical model, does include the resources required by a commitment in its definition. Jennings (1993) specifies and implements an architecture based on the model of joint responsibility in which agents can make both individual and joint commitments. Agents also have an explicit social convention which specifies under what circumstances commitments can be reneged upon and how to act towards others in such cases.

As well as providing the basis for agent architectures, commitments have also been used as key components of DAI programming languages. Shoham (1993) uses commitments in his work on specifying a computational framework for agent programming. Agent-Oriented Programming is a specialisation of object oriented programming in which the mental state of an agent (object) consists of the precisely defined components: beliefs, commitments and capabilities. Commitment is a primitive feature of the programming language and has the form that at a particular time t, agent A is committed to agent B about G where G can be a belief or an action. In this formalism, commitment is viewed as an inherently social phenomena with commitments made to the self being a special case (agents A and B being the same entity)[7].

Bond (1989) also outlines a language for social problem solvers in which agents are programmed in terms of their commitments to one another. Agents are described in terms of: the actions they will perform and the resources they will need; the beliefs they hold; their expectations about the actions of others and the resources they will supply and, finally, the resources they will supply to others by fulfilling their obligations. Commitment is represented as a logical literal which can be a goal, belief or action and which has an associated specification of the resources it will utilise.

---

[7]. Fikes (1982) also uses the notion of making commitments to others as the basis of his framework for describing cooperative work in informal domains (such as an office).

In both languages commitments are primitive concepts and there is no confusion with notions of reneging upon obligations. Conventions can be specified separately using other primitives of the agent language and by detailing various properties for the persistence of commitments over time. However it is doubtful whether joint commitments could be encoded in either language as it stands at present - they would need to be defined as a new language primitive.

3 The Commitment and Convention Model of Coordination

This section describes a new unifying model of coordination which has the concepts of (joint) commitments and (social) conventions at its core. This model distills, synthesizes and clarifies many of the important features of the extant models of coordination presented in the previous section. Using a distributed goal search characterisation of DAI (section 3.1), the "Centrality of Commitments and Conventions Hypothesis" is explained and argued for:

Centrality of Commitments & Conventions Hypothesis

All coordination mechanisms can ultimately be reduced to (joint) commitments and their associated (social) conventions.

3.1 Distributed AI as Distributed Goal Search

Several authors have recently characterised DAI as a form of distributed goal search with multiple loci of control (Durfee and Montgomery, 1991; Gasser, 1992b; Lesser, 1991). Adopting Lesser's basic formalism, the actions of $Agent_1$ and $Agent_2$ in respectively solving goals G^1_0 and G^2_0 can be expressed as a classical AND/OR goal structure search[8] (figure 2). The classical structure has been augmented to include the representation of interdependencies between goals because these are the key to coordination in DAI systems. The resources needed to solve primitive goals (leaf

[8] Figure 2 represents a situation in which each individual has its own goals, but to achieve them it must interact with others. To represent a DAI system in which all community members pursue a common goal, there would be a single root node corresponding to the shared objective.

nodes) are also shown. Interdependencies can exist between high level sibling goals, such as G^1_1 and G^1_2, or they can be more distant in the goal structure (eg between $G^1_{1,1}$ and $G^2_{p,2}$). In the latter case, G^1_1 and G^2_p become interacting goals if $G^1_{1,1}$ is used to solve G^1_1. Indirect dependencies exist between goals through shared resources (eg $G^1_{m,1,2}$ and $G^2_{p,2,2}$ through resource d^1_j). Resource dependencies can be removed simply by providing more of the resource in question; dependencies between goals, on the otherhand, cannot be circumvented as they are a logical consequence of the community's environment. In all other aspects the two types of dependence are the same.

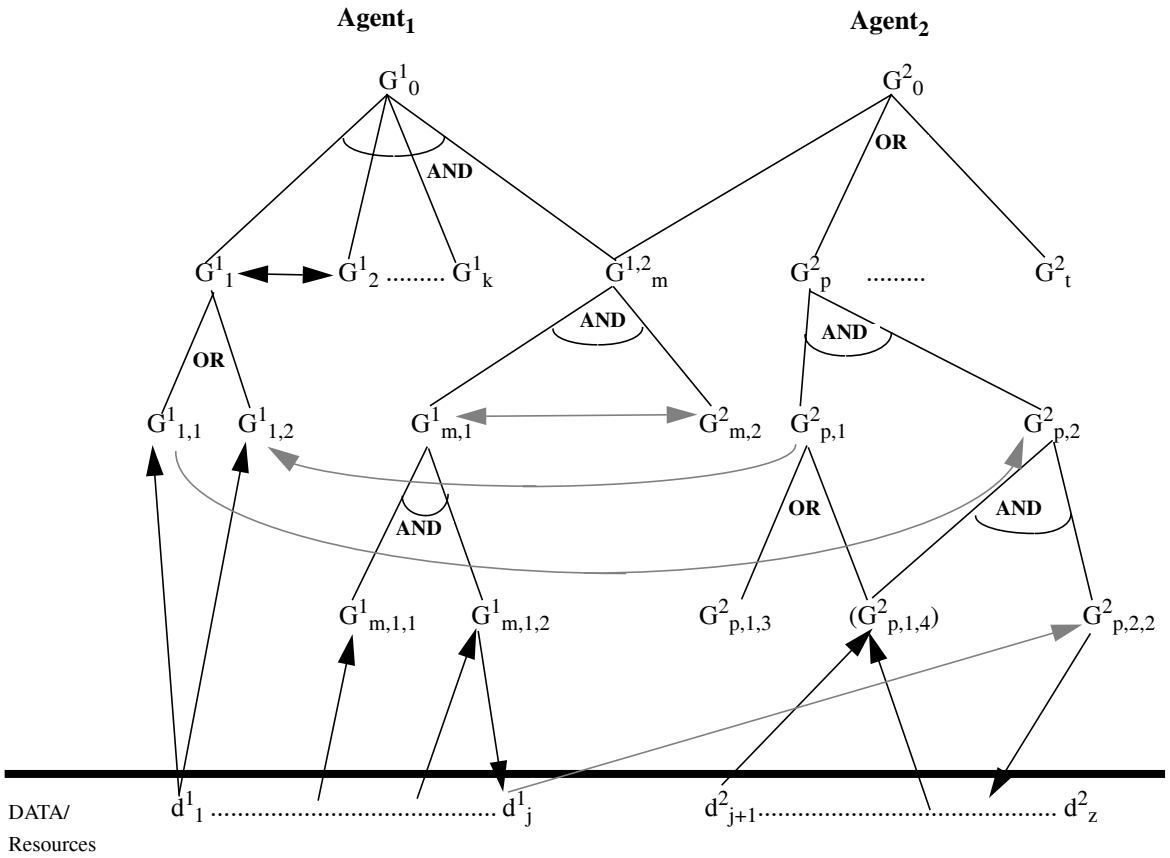

Figure 2: A distributed goal search tree involving $Agent_1$ and $Agent_2$. The dotted arrows indicate interdependencies between goals and data in different agents, solid arrows dependencies within an agent. The superscripts associated with goals and data indicate the agent which contains them.

Interdependencies can be classified along two orthogonal dimensions: whether they are weak or strong and whether they are uni-directional or bi-directional. S*trong dependencies* must be

satisfied if the dependent goal is to succeed, *weak dependencies* facilitate problem solving but need not be fulfilled for the dependent goal to complete. An example of a strong dependency is where the output of a goal (G) is a mandatory input (I) for the dependent goal (DG) and that G is the only source of I in the community; a weak dependency would exist if there was more than one source for I or if I was an optional input for DG. A *uni-directional* dependency (written $G^1_{1,1}$ ---> $G^2_{p,2}$) means that $agent_2$'s goal $G_{p,2}$ is dependent (either strongly or weakly) on $agent_1$'s goal $G_{1,1}$, but $G^1_{1,1}$ is unaffected by $G^2_{p,2,}$; with *bi-directional* dependencies (written $G^1_{m,1}$ <---> $G^2_{m,2}$) the goals of both agents are affected. The providing of information I by goal G for goal DG, above, is an example of a uni-directional dependence (G ---> DG); a bi-directional dependence occurs, for example, when two actions need to be performed simultaneously.

For this work it was necessary to extend Lesser's graph formalism to allow joint goals (eg $G^{1,2}_m$). Joint goals represent inherently social actions or objectives which a group of agents have decided to solve as a team. They provide the glue to bind individuals' actions into a cohesive whole and must ultimately give rise to individual goals as only individual agents have the ability to act (perform leaf node tasks). Joint goals can be in the mind of each individual who is acting as part of the collective, implying that everything necessary for team behaviour can be possessed by individual agents even though the aim makes reference to the collective. Thus the joint goal $G^{1,2}_m$ is internalised within $Agent_1$ and $Agent_2$ and results in $Agent_1$ performing $G^1_{m,1}$ and $Agent_2$ performing $G^2_{m,2}$. Such joint action requires a shared objective which the group wishes to achieve and a recognition that they want to achieve it in a collaborative manner. So in a collaborative lift, for example, all team members must want to lift the object and they must want to do so as part of a group effort. The second component of the definition is important because it distinguishes between identical and parallel goals (Conte *et al.*, 1990). For instance if both x and y have the goal to cook spaghetti then their goals are identical; but if both agents have the goal to eat spaghetti (i.e. x has the goal that x eats spaghetti and y has the goal that y eats spaghetti) they merely have parallel goals. These two goal types result in different forms of social action - identical goals can give rise to joint goals if the two agents decide to work as a team, whereas

parallel goals give rise to competition and will certainly not result in joint action.

Lesser (1991) makes the following general observations about the graph formalism. The entire goal structure need not be fully elaborated in order for problem solving to begin, it may be constructed as problem solving progresses. Actually developing the graph can be a complex social activity involving negotiation, persuasion and the resolution of conflicts or it may be undertaken centrally by one agent. Construction can involve a top-down elaboration based on the higher-level goals, a bottom up process driven by the data, or a mixture of the two. Finally, the formalism says nothing about whether the structure is statically defined or dynamically evolves from a composite view of the current, local goal structures of the individual agents.

3.2 Coordination as Control over Distributed Goal Search

Formulating a DAI system as a distributed goal search problem allows the activities which may require social interaction to be clearly identified, these include: (i) defining the goal graph (including identification and classification of interdependencies); (ii) assigning particular regions of the graph to appropriate agents; (iii) controlling decisions about which areas of the graph to explore; (iv) executing (traversing) the goal structure; (v) ensuring that successful traversal of the search space is reported. Some of these activities may be done in a collaborative fashion and some may be done by one individual. Determining which approach is adopted for each of the various phases is a matter of system design. It will depend upon the nature of the domain (eg in applications in which agents have distinct expertise, assignment of goals simply becomes a matter of identifying the individual capable of performing the activity), the type of agents which are included in the community (eg with autonomous agents, the global search space is given by the union of the local search spaces and each agent works on its own local goals) and the desired solution characteristics (eg to increase the likelihood of an important result being produced, the same area of the search space may be redundantly assigned to multiple agents, whereas if the desire is to optimise agent usage then such an arrangement is inefficient).

Here consideration of the coordination process is restricted to deciding which areas of the graph to explore, actually executing the goal structure and ensuring successful traversal of the goal graph is reported. Assuming there are interdependencies between the goals, or amongst the resource requirements, of the different agents, coordination is desirable (sometimes essential) if the community is to act in a coherent manner.

The nature of this dependency is the critical determinant of the type of coordination. For example if $Agent_1$ knows that $G^2_{p,2,2}$ requires resource d^1_j before it can start (strong dependency, uni-directional), then it may decide to execute $G^1_{m,1,2}$ (to produce the necessary resource) before $G^1_{m,1,1}$ if there is no other information distinguishing between these two alternatives. Secondly the relationship between $G^1_{m,1}$ and $G^2_{m,2}$ may stipulate that both actions need to be performed simultaneously (strong dependency, bi-directional) in which case the two agents need to reach an agreement about the respective execution times[9]. Finally, if $Agent_1$ chose $G^1_{1,1}$ as a means of satisfying G^1_1 the result of this task may provide valuable information (weak dependency, uni-directional) which $Agent_2$ could use when solving $G^2_{p,2}$ (eg it may provide a partial result which enables $G^2_{p,2}$ to be significantly shorter). Knowing this, $Agent_2$ may start with $G^2_{p,1}$.

3.3 Commitments

The following points represent the important intuitions present in the extant models of coordination. The term “commitment” means a pledge or promise. Agents can make pledges both about actions and beliefs and these pledges can either be about the future or the past. Thus agent A can commit itself to play cricket tomorrow (object of commitment = action, time = future) and agent B can commit itself to believe a particular version of events about the reasons for the start of World War I (object of commitment = belief, time = past). For the purposes of coordination, however, the most important commitments are related to future actions. No fundamental

9. Other types of temporal relationship, such as BEFORE, DURING and OVERLAPS (Allen, 1984), can also be modelled in this manner.

difference between pledges which are internalised within an agent (eg I will lose 12 pounds in weight) and those which are made to a second party (eg I will fix your car for you) are assumed. Commitments may be conditional - for example A will play cricket tomorrow if the weather is sunny. Finally, pledging to undertake an activity involves an associated commitment about the resources required to carry out that action. So if A pledges to play cricket tomorrow, then it is also devoting its resources of time and energy to this activity.

If an agent commits itself to perform a particular action then, provided that its circumstances do not change, it will endeavour to honour that pledge. This obligation constrains an agent's subsequent decisions about undertaking fresh activities since it knows that sufficient resources must be reserved to honour existing commitments. If an agent had infinite resources which could be freely allocated to any permutation of its commitments then there would be no such restriction. However as most resources are finite, and also because there are often constraints imposed by the environment, an agent is limited in the number and type of commitments it can make[10]. For this reason, an agent's commitments should, as far as it is aware, be both internally consistent and consistent with its beliefs (Bratman, 1990). The former means that an individual's commitments should not conflict with one another - for example an agent should not pledge to simultaneously perform two goals which both require the same non-shareable resource. The latter means that if an agent's intended actions are executed in a world in which its beliefs are true, the desired state of affairs will ensue. This accounts for the fact that an agent's beliefs will, in most cases, be both partial and imprecise - meaning commitments may be made on false premises and therefore turn out to be unachievable or inappropriate.

Joint commitments have all the aforementioned properties of individual commitments, but have the additional constraint that they involve more than one agent[11]. This means the overall

[10.] Commitments provide a "filter of admissability" (Bratman, 1984), agents should not commit themselves to something which will conflict with or endanger their existing commitments without good cause.

[11.] A joint commitment involving one agent is equivalent to an individual commitment.

state of the joint commitment is distributed; with individual commitments the agent which has made the pledge is aware of its status since it forms part of its mental state. So, for example, the state of joint commitment $G^{1,2}_{m}$ is distributed between $Agent_1$ in its processing of $G^{1}_{m,1}$ and $Agent_2$ through its processing of $G^{2}_{m,2}$. Ideally all team members should have access to a shared mental state related to the joint commitment as this would ensure they all simultaneously have the same experiences and beliefs and also that there can be no divergence amongst the group's members. However since group activity is undertaken by individuals and not the team *en mass*, it is the individuals who have first exposure to events related to the joint commitment. Thus a shared mental state is impossible unless all the agents possess a single common structure which records all of their beliefs about the joint commitment (i.e. agents cannot have any local or private beliefs about the joint action). For example in a team search, if one agent satisfies the group's objective and finds the target item then at that precise instant in time it is the only one who knows that the joint commitment has been fulfilled. This agent may subsequently inform the others of its achievements, meaning they all share a common perspective once more, however in the meantime different members of the group have diverged in their beliefs about the joint commitment. In all other respects, the difference between the two types of commitment is merely quantitative (eg a joint commitment is, in general, likely to contain more interdependent goals, but the types of relationships will be identical to those which can be found between individual commitments).

3.4 Conventions

An agent should honour its commitments provided that its circumstances do not change. However in most realistic scenarios, agents are situated in time-varying contexts - the external world may change, the agent may become aware of new information, another agent may attempt to interact with it, and so on. Therefore in many cases an agent's beliefs will alter between the making of a commitment and it actually performing the associated processing - in fact, the longer the time between these two events the greater the likelihood of a change occurring. In some instances, these changes will leave the agent's commitment unaffected, however in other cases commitments

may need to be reviewed. For example if agent A is informed that the first customer at a new garage opening tomorrow will receive a Ferrari then it may indeed revise its commitments about playing cricket. Commitments should, therefore, be relatively stable over time (otherwise there is little point in making them), but they should not be irrevocable.

To operate successfully and intelligently, agents need general policies for governing the reconsideration of their commitments. These *conventions* describe circumstances under which an agent should reconsider its commitments and indicate the appropriate course of action to either retain, rectify or abandon the commitment. When specifying conventions a balance needs to be reached between constantly reconsidering all commitments (which will enable the agent to respond rapidly to changing circumstances, but means it will spend a significant percentage of its time reasoning about action rather than actually carrying out useful tasks) and never reconsidering commitments (which means agents spend most of their time acting, but what they are actually doing many not be particularly relevant in the light of subsequent changes). Kinny and Georgeff (1991) carried out a series of experiments in which different conventions were examined in environments which exhibited different rates of change. In all cases it was found that the "bold" agents (those which never reconsidered their plans) performed better than the "normal" agents (those which are slightly more open to reconsideration) which were better than the "cautious" agents (those which were prone to reconsideration). However in rapidly changing and uncertain environments the utility of a relatively sophisticated convention is significantly increased. Indeed empirical evaluation has shown that in such circumstances conventions play a pivotal role in ensuring the community acts in a coherent manner (Jennings and Mamdani, 1992).

Both the list of situations under which commitments should be reassessed and the actions which should be taken in such circumstances can be empty. So an agent could remain permanently committed to a goal even if it has been achieved and an agent could take no actions as a result of changes in its circumstances. The action part is particularly useful in multiple agent environments because it specifies how any interdependencies with other commitments should be

CONVENTION: Limited-Bandwidth

REASONS FOR RE-ASSESSING COMMITMENT:

- COMMITMENT SATISFIED
- COMMITMENT UNATTAINABLE
- MOTIVATION FOR COMMITMENT NO LONGER PRESENT

ACTIONS:

```
R1: IF   COMMITMENT SATISFIED OR
         COMMITMENT UNATTAINABLE OR
         MOTIVATION FOR COMMITMENT NO LONGER PRESENT
    THEN DROP COMMITMENT

R2: IF COMMITMENT SATISFIED
    THEN INFORM ALL RELATED COMMITMENTS

R3: IF COMMITMENT DROPPED BECAUSE UNATTAINABLE OR MOTIVATION NOT PRESENT
    THEN INFORM ALL STRONGLY RELATED COMMITMENTS

R4: IF COMMITMENT DROPPED BECAUSE UNATTAINABLE OR MOTIVATION NOT PRESENT
         AND COMMUNICATION RESOURCES NOT OVERBURDENED
    THEN INFORM ALL WEAKLY RELATED COMMITMENTS
```

Figure 3: Sample Convention

dealt with (see R2, R3 and R4 in figure 3), although it may also be used for dealing with purely internal matters related to the dropping of commitments (see figure 1). Figure 3 gives an example convention in which the reasons for reassessment are based on Cohen and Levesque's formalism (see section 2.1) and the actions are designed for operating in a multi-agent community in which the communication bandwidth is limited.

An agent may have several conventions at its disposal. Although there is no intrinsic difference between a convention for purely local goals (see figure 1) and one which has external dependencies (as in figure 3), the overall coherence of the community will be improved if some minimum reporting actions are included (Jennings and Mamdani, 1992). Therefore an agent may have some conventions for dealing with purely local commitments, some for dealing with commitments which are weakly related to others and some for handling strong dependencies. An agent must have precisely one convention for each of its active commitments, although it may use different conventions for different commitments.

For inter-agent dependencies, the participants should, ideally, be mutually aware of the convention which governs their interaction. Such awareness is needed if the agents are to minimise the uncertainty in their collaboration and maximise the benefit of the coordination process. Thus, for example, if $Agent_2$ must have resource d^1_j to perform $G^2_{p,2,2}$ then it will ask $Agent_1$ to make this resource available. However merely asking for d^1_j to be produced is not sufficient, because $Agent_2$ also wants to be informed when it is available. To ensure the necessary dissemination occurs, $Agent_2$ must also request that the resource is produced using an appropriate convention (eg the Limited-Bandwidth convention of figure 3). Whether $Agent_1$ accepts this convention proposal will depend on its preferences and the relative authority relationship of the two agents. If the proposal is acceptable, or $Agent_2$ can force $Agent_1$ to use it, then the convention will be adopted. If the proposal is unacceptable, then the two agents will have to enter a negotiation phase to decide upon an acceptable solution. Alternatively, rather than having to determine the convention for each and every interdependent goal at runtime, which will significantly slow down processing, the system designer may stipulate that when two agents interact they will always use a particular convention. He may even specify that the whole community must use a particular convention for all their interactions.

Although agents engaged in a joint commitment cannot have a shared mental state, it is important that relevant information pertaining to their commitment is disseminated at the earliest possible opportunity. However agents should not broadcast information about their commitments every time they change as this will overburden the communication resources and needlessly distract the recipients. Rather they should only inform those agents who are likely to be affected by their change. Such interchange aims to provide an approximation to a common state and tries to minimise the effects of distribution whilst not overburdening the communication channels. The two fundamental pieces of information which must be shared are: (i) the status of the commitment to the shared objective; (ii) the status of the commitment to the given team framework[12]. If an agent's beliefs about either of these key issues changes, then it is part of the "cooperativeness" inherent in joint goals that all team members are informed. Many joint actions depend upon the

BASIC SOCIAL CONVENTION

REASONS FOR RE-ASSESSING COMMITMENT:

- STATUS OF COMMITMENT TO SHARED OBJECTIVE CHANGES
- STATUS OF COMMITMENT TO REACHING SHARED OBJECTIVE IN PRESENT TEAM CONTEXT CHANGES
- STATUS OF JOINT COMMITMENT OF A TEAM MEMBER CHANGES

ACTIONS:

R1: IF STATUS OF COMMITMENT TO SHARED OBJECTIVE CHANGES OR
STATUS OF COMMITEMENT TO PRESENT TEAM CONTEXT CHANGES
THEN INFORM ALL OTHER TEAM MEMBERS OF CHANGE

R2: IF STATUS OF JOINT COMMITMENT OF A TEAM MEMBER CHANGES
THEN DETERMINE WHETHER JOINT COMMITMENT STILL VIABLE

Figure 4: Minimum Convention for Joint Commitments

participation of all their team members, therefore a change of commitment from one participant can jeopardise the whole group's efforts. Hence if an agent comes to believe that a fellow team member is no longer jointly committed, it also needs to reassess its position with respect to the shared objective. These basic assumptions are encoded in a convention which represents the minimum state of affairs for joint commitments (figure 4) - this convention is similar to the persistence of a weak achievement goal in Cohen and Levesque's formalism for joint goals (see section 2.2.1). Thus whereas any conventions can be used for individual commitments, including the empty one, joint commitments require each team member to adhere to the minimum social convention. This requirement is the sole distinguishing characteristic between individual agents which have highly interrelated commitments and a team of agents which have a joint commitment.

In certain applications it may be desirable to have more sophisticated social conventions

[12.] The second stipulation covers the situation in which an agent, which is initially jointly committed to a collaborative act, decides to leave the team but continues to pursue the shared objective in an individualistic manner. In this case the agent can no longer be said to be jointly committed, since it will follow its own solution path without consideration of its affects on the others who remain in the original team.

JOINT RESPONSIBILITY SOCIAL CONVENTION

INHERIT: BASIC SOCIAL CONVENTION

REASONS FOR RE-ASSESSING COMMITMENT:

- SHARED OBJECTIVE IS MET
- SHARED OBJECTIVE WILL NEVER BE MET
- MOTIVATION FOR SHARED OBJECTIVE IS NO LONGER PRESENT
- AGREED PLAN WILL NOT ACHIEVE DESIRED RESULTS
- AGREED PLAN CANNOT BE EXECUTED
- AGREED PLAN HAS NOT BEEN EXECUTED PROPERLY

ACTIONS:

R1: IF SHARED OBJECTIVE IS MET OR
SHARED OBJECTIVE WILL NEVER BE MET OR
MOTIVATION FOR SHARED OBJECTIVE IS NO LONGER PRESENT
THEN DROP JOINT COMMITMENT TO SHARED OBJECTIVE & TO AGREED PLAN

R2: IF AGREED PLAN WILL NOT ACHIEVE DESIRED RESULTS OR
AGREED PLAN CANNOT BE EXECUTED OR
AGREED PLAN HAS NOT BEEN EXECUTED PROPERLY
THEN DROP JOINT COMMITMENT TO AGREED PLAN

R3: IF DROP JOINT COMMITMENT TO AGREED PLAN AND
CAN RE-PLAN USING SAME AGENTS
THEN DEVELOP AND JOINTLY COMMIT TO NEW PLAN

R4: IF DROP JOINT COMMITMENT TO AGREED PLAN AND
CANNOT RE-PLAN USING SAME AGENTS AND
CAN DEVELOP NEW PLAN USING DIFFERENT TEAM
THEN DROP JOINT COMMITMENT TO EXISTING TEAM & JOINTLY COMMIT TO
NEW TEAM

R5: IF CANNOT DEVELOP NEW COMMON PLAN
THEN DROP JOINT COMMITMENT TO SHARED OBJECTIVE & TO AGREED PLAN

Figure 5: Joint Responsibility Convention for Joint Commitments

which build upon the basic one. For example when agents are situated in environments in which they possess neither complete nor correct beliefs about their world or other agents, have changeable goals and fallible actions and are subject to interruption from external events, reconsideration of commitments and decisions about subsequent actions becomes a primary consideration. In these complex and dynamic environments it is difficult to ensure that a group's behaviour remains coordinated, because initial assumptions and subsequent deductions may be

incorrect or inappropriate. Joint responsibility (section 2.2.2) is an example of a coordination model whose convention was designed to operate in just such situations (figure 5).

3.5 Commitments, Conventions and Coordination

With respect to coordinating the behaviour of multiple agents, the most important feature of commitments is that they enable individuals to make assumptions about the actions of other community members. They provide a degree of predictability to counteract the uncertainty caused by the distribution of control. So for the joint goal $G^{1,2}_{m}$, $Agent_2$ can carry out $G^{2}_{m,2}$ in the knowledge that $Agent_1$ is probably performing $G^{1}_{m,1}$ and that if it is not, then it will at least be trying to inform it of this change (because of the basic social convention). Without this assurance there would be no point in $Agent_2$ even starting $G^{2}_{m,2}$ since it is only carrying out this activity to achieve the joint goal and the joint goal requires both sub-goals to be fulfilled. Thus each agent is only carrying out its action because it believes that the other is also doing its bit. Commitments also enable agents to reason about the activities of others in deciding how to adapt their local problem solving behaviour to benefit from social interactions. For example if $Agent_2$ knew that $Agent_1$ was committed to $G^{1}_{1,1}$, whose outcome enables it to improve its solution of $G^{2}_{p,2}$, then it may decide to wait for this information and process $G^{2}_{p,1}$ in the meantime.

Commitments can be made at many different levels and have correspondingly varied time horizons. When $Agent_1$ commits itself to perform G^{1}_{0} this will invariably be a high level objective (eg diagnose faults in an electricity network) to which the agent will probably remain committed for some considerable amount of time. The leaf nodes, on the otherhand, will involve fairly specific courses of action (eg see if there is a fault in low voltage $line_1$) and have a much shorter duration.

Generally the greater the degree of accuracy to which an agent knows its acquaintances commitments, the more detailed its predictions can be and so the more coherently the community will behave. However it is not always desirable to transmit all of the low-level details about

commitments - rather agents should communicate at a sufficiently detailed level to promote satisfactory coordination, but at a sufficiently abstract level to ensure agents retain sufficient flexibility in achieving their objectives in an uncertain environment. For example knowing that $Agent_2$ is committed to G^2_0, gives no indication whether $G^2_{p,1}$ will be performed. However knowing that $Agent_2$ is committed to G^2_p means that it is possible to predict that $G^2_{p,1}$ will indeed be performed and that $Agent_1$ could delay its processing of $G^1_{1,2}$ to benefit from the weak, uni-directional dependence. If $Agent_2$ communicated a more detailed description of its intentions, eg that it will perform $G^2_{p,1,3}$, then this information is of no additional benefit to $Agent_1$ since it is not dependent on how $G^2_{p,1}$ is achieved. Not sending details of how $G^2_{p,1}$ will be achieved also leaves $Agent_2$ unconstrained as to whether it will use $G^2_{p,1,3}$ or $G^2_{p,1,4}$.

3.5.1 Commitments involving Goals and Sub-Goals

If there is an AND relationship being undertaken by a cooperating group, then all commitments to the sub-goals must be honoured if the parent goal is to succeed. If just one agent reneges, then the other agents' activities are doomed in their present form. Therefore it is only the belief that others will honour their commitments, which makes it rational for an agent to carry out its part. Without this confidence no agent would carry out its individual processing, since achievement of the sub-goals in isolation is unlikely to bring it any benefits. Hence AND goals cannot be achieved by cooperative problem solving without the notion of commitment.

If a goal has a number of sub-goals, organised in an OR relationship, which are each assigned to different agents who all carry out their activities in parallel, then failure of one agent to fulfill its commitment will not jeopardise achievement of the parent goal. However if all the agents renege upon their obligations then the parent goal will not be achieved; thus the commitment of at least one agent must be guaranteed. If agents coordinate their activities more closely and arrange for only one of the alternate sub-goals to be carried out at any one time (to avoid needless duplication) then commitment failure can have serious consequences for the community's overall level of coherence. For example if a team agrees that $Agent_1$ will carry out the subgoal which

fulfills the parent goal G, then the remaining agents can continue with their processing and can make subsequent commitments based on the fact that G will be achieved and that they do not have to expend resources towards this end. However if $Agent_1$ does not fulfill its pledge, then provided that G is still desired, one of the other agents will have to carry out some unexpected processing activity. This additional work may conflict with commitments which the agent has subsequently made and may result in it having to delay or even abandon some of them because additional resources are unexpectedly required to achieve G. Such delays may then have a knock-on effect to other agents, causing the community to operate ineffectively and requiring it to undertake a significant amount of replanning.

3.5.2 Commitments involving Dependencies between Goals

Consider the situation in which $Agent_1$ and $Agent_2$ have the respective interrelated goals G^1_1 and G^2_1. If there is a strong bi-directional dependence then both agents must honour their commitments otherwise neither of them will be able to achieve their objectives. If the relation is strong but uni-directional (G^1_1 ---> G^2_1), then failure of $Agent_1$ to honour its commitment means that $Agent_2$ will be unable to achieve G^2_1 and it will either have to find an alternative path for achieving the parent goal or abandon it completely.

With weak dependencies the agents involved may still be able to proceed but possibly in a suboptimal manner. For example an agent may have delayed processing an action on the premise that an acquaintance will provide it with sufficient information to significantly speed up its problem solving. If this information is no longer forthcoming, because the acquaintance changed its commitments, then the agent has wasted potentially useful processing time. As another example, an agent may select a certain path in the belief that information which will be provided through a weak dependency will make this path less expensive than its alternatives. But if the agent providing the information reneges upon its commitment and the information is not forthcoming then the chosen path may be suboptimal.

If the relationship is bi-directional and one agent fails to fulfill its pledge, the agent which is still committed to its side of the bargain may be adversely affected since it chose to undertake the goal believing that it would be able to profit from the commitment of the other agent. In the uni-directional case (G^1_1 ---> G^2_1) if $Agent_1$ changes its mind, $Agent_2$'s processing of G^2_1 will be adversely affected for the reason described above.

In both the weak and the strong uni-directional cases, if $Agent_2$ drops its commitment to G^2_1 then this may even have a detrimental affect on $Agent_1$. This is the case if $Agent_1$ chose G^1_1, even though it was locally suboptimal, because the net utility to the community of the performance of the pair $\{G^1_1, G^2_1\}$ whilst satisfying the specified relationship was higher than if $Agent_1$ chose an alternative to G^1_1 and $Agent_2$ chose G^2_1. However the potential benefits of $Agent_1$'s sacrifice were not observed because $Agent_2$ failed to carry out its pledge about G^2_1.

3.5.3 Conventions and Goal Interrelationships

Conventions which report changes in commitments to dependent agents are especially important when there is an AND relationship between a goal and its constituent sub-goals. Without adequate information dissemination, the other agents will remain committed to performing their sub-goals which will not satisfy their original purpose of fulfilling the parent goal. This is also the case for OR relationships in which only one of the sub-goals is active at any one time.

In terms of interagent dependencies (G^1_1 --> G^2_1), reports of changes in the status of commitments are essential if the relationship is strong and bi-directional and if $Agent_1$ reneges on its goal and the relation is strong and uni-directional. In all other cases reports on changes of commitments are desirable in that they may enable the agent which is still committed to reassess its position. This may result in it choosing a different path through the graph - either because it is freed from the constraint of having to honour the relationship or because it can no longer benefit from the interaction with the agent which is no longer committed.

4 Coordination without Commitments and Conventions?

In addition to the models of section two, a number of other approaches to coordination have been developed which do not contain explicit references to either commitments or conventions. Three such mechanisms which are common in DAI are: organisational structuring, exchanging meta-level information and multi-agent planning. Each of these approaches is examined in turn; there is a brief statement about how it facilitates the coordination of behaviour, what its main characteristics are and how it can be reformulated into the Centrality of Commitments and Conventions Hypothesis. Choosing between coordination mechanisms for a particular application is a matter of system design - there is no universally best approach. Indeed several multi-agent systems embody more than one approach because of the differing time horizons, level of detail and communication requirements (eg Durfee and Montgomery's (1991) work on coordination in a hierarchical behaviour space).

4.1 Organisational Structures

In the context of multi-agent systems, an organisational structure can be viewed as a pattern of information and control relationships which exist between individuals within the community. Control relationships can be hierarchical, heterarchical or flat and are responsible for designating the relative authority of the agents and for shaping the types of social interaction which can occur. Organisational structure can also be used as a high-level specification of the distribution of problem solving capabilities amongst community members (Durfee *et al.*, 1989). For example when building a community of agents for diagnosing faults in an electricity network (Aarnts *et al.*, 1991; Cockburn *et al.*, 1992), the system designer may specify a functional organisation (agent_1 works on high voltage faults, while agent_2 works at the low voltage level) or a spatial organisation (agent_1 deals with all types of faults in region_1, agent_2 with all types of faults in region_2). In a spatial distribution in which there are overlaps, authority relationships determine how redundances are avoided - in a hierarchy, high level nodes inform the lower level ones of the activities they are to pursue, whereas in a flat structure this process will only be achievable

through direct negotiation between the parties concerned.

Organisational structures give general, long term information about the relationships between agents. When viewed as a distribution of capabilities they specify which actions an individual will undertake and provide a means of dividing up the search space without having to go into detail about the particular sub-trees. Other authors have followed this basic approach using different terminology. Singh (1990) employs the notion of strategies to provide an abstract specification of the behaviour of an agent or a group and Werner uses roles for describing expectations about individual behaviour (see section 2.1). Shoham and Tennenholtz (1992) propose a more detailed organisational form which they term a "social law". In this approach the society adopts a set of laws (eg road traffic rules) which specify how individuals should behave. Each programmer is then committed to obeying these laws when building his individual agent and can assume that all the others will as well.

Organisational structures aid the process of coordination by providing a high-level view of how the community solves its problems and also by identifying the role of each individual. For example the Distributed Vehicle Monitoring Testbed (Lesser and Corkill, 1983) simulates a spatially organised community of agents which performs a distributed interpretation to track vehicles moving amongst them. Each agent decides which areas of the search space to explore based upon its current local view, but uses organisational knowledge about its problem solving role in the community and the roles of others to guide its decisions so that it is a more effective participant in the community. With this approach coordination consists of two concurrent activities: the construction and maintenance of a community wide organisational structure and the continuous elaboration of this structure into precise activities using the local knowledge and control capabilities of each agent. In this context the organisation is specified as a set of "interest areas" and as a set of ratings of priority. The former indicate what, when and to whom information should be sent, the latter indicate how to evaluate the importance of processing different types of goals. Authority relationships indicate the relative priorities which should be attached to

processing externally generated goals versus local goals.

When an agent undertakes a particular role within an organisation it is, in fact, making a high-level commitment about the types of activity it will pursue. For instance in the electricity management scenario, if $agent_1$ undertakes the role of diagnosing high-voltage faults, other agents will expect it to undertake work in this area. They will make subsequent decisions in their local problem solving based on the assumption that $agent_1$ will indeed be dealing with all the faults on the high voltage network.

Although they are relatively long-term structures, it has been shown that different organisations are appropriate for different problem situations and performance requirements (Malone, 1987). Hence as a situation evolves, the community may need to periodically reassess its structure to determine whether it is still appropriate or whether a rearrangement would be beneficial - see the work of Ishida *et al.* (1990) for an illustration of the dynamic reorganisation of a group of cooperating agents in response to changes in the environment. In the electricity management scenario, for example, the community may decide that it is best to replace the agent carrying out high-voltage diagnosis with several spatially distributed agents so that the load and the reliance upon any one individual is reduced. This evaluation corresponds to a convention for the organisational structure.

4.2 Meta-Level Information Exchange

Meta-level information is control level information about the current priorities and focus of a problem solver (Gasser, 1992b) - it indicates approximate regions of the search space on which agents will concentrate most of their efforts. For example in the functionally distributed electricity management scenario described in section 4.1, $agent_1$ may indicate that it believes the most important fault is in $region_1$. Upon receiving this information $agent_2$, who is working on the high-voltage network, may also decide to concentrate its efforts on this region to determine whether the fault being experienced on the low voltage network is in fact a manifestation of a problem with

the high voltage system (eg no supply is getting through).

Durfee (1988) has developed a meta-level information exchange approach to coordination called Partial Global Planning in which agents build and share local plans as a means of identifying potential improvements to coordination[13]. These partial global plans (PGPs) are exchanged by agents as a means of building representations of their acquaintances' activities - they indicate which goals will be pursued, in what order, what results will be achieved and how long each goal is likely to take. Individual community members then use a model of themselves and a representation of their acquaintances to identify when agents have PGPs whose objectives are part of some larger community effort. If such complimentary activities are detected the related PGPs are combined into a single, larger PGP which provides a more complete view of the group's activity. Agents can then revise their PGPs to reflect the new position and may consequently decide to alter their local plans to better utilise the community's resources. For example a PGP could indicate that a partial solution to be formed by one agent provides useful predictive information for an acquaintance. This expectation and the transmission of the partial solution would then be explicitly represented in the PGP, resulting in a plan to use information resources more effectively. As a second example, an agent may survey its current view of community-wide PGPs and identify acquaintances who are being under utilised, whilst there are others who are overburdened. By modifying its PGPs, the agent could propose how the community could transfer subproblems so as to work better as a team.

Meta-level information exchange is a medium term source of knowledge about an agent's commitments - shorter than organisation structures but longer than multi-agent planning approaches. It enhances coordination only to the degree to which it is accurate - indeed inaccurate information may be more detrimental than no information at all. Again it can be seen that once an

[13]. This approach differs from standard multi-agent planning (section 4.3) in that agents might never reach mutual agreements about their multi-agent commitments because the partial plans can change so fluidly and also because it takes time to propagate changes.

agent indicates it will work in a particular region of the search space, it is important to honour that commitment - failure to do so will result in misleading information being spread around the network and incoherent problem solving. However, as with the other approaches, commitments should not be irrevocable; some form of convention is needed for monitoring their progression. With the PGP approach, for example, agents often altered their local plans, either because new tasks arrived or because actions took longer than expected, and so their commitments needed to be updated. However if agents informed each other of every minor change in their commitments, it could cause a chain-reaction which spreads throughout the system. Therefore agents adopted an implicit convention in which they informed their acquaintances only when the deviations were deemed significant.

4.3 Multi-Agent Planning

With this approach to coordination, agents usually form a multi-agent plan which specifies all of their future actions and interactions with respect to achievement of a particular objective. It details, before execution commences, the areas of the search space that will be traversed and the route to be taken at each decision point for each agent involved in the activity. Multi-agent plans are typically built to avoid inconsistent or conflicting actions, particularly with respect to consumption of scarce resources.

Multi-agent planning differs from organisational structuring and meta-level information exchange in terms of the level of detail to which it specifies every agent's activities. Agents know in advance exactly what actions they will take, what actions their acquaintances will take and what interactions will occur. By requiring such a complete specification of behaviour, the plans can only realistically have a short time horizon because of problems with the unpredictability and dynamicity of events in the environment. As plan construction has to take into account all the possible choice points the agent would have reached, without the benefit of constraining information from actual execution, this approach often requires substantially more computational and communication resource than the other two mechanisms.

There are two basic approaches to multi-agent planning: centralised and distributed. Georgeff (1983) developed a system in which the plans of individual agents were developed separately and then sent to a central coordinator who analysed them to identify potential interactions. The coordinator identified interactions which could cause conflicts and grouped together sequences of unsafe situations to create critical regions. Finally it inserted communication commands into the individual plans so that agents synchronise their activities appropriately. Cammarata *et al.* (1983) have devised a centralised multi-agent planning system for air traffic control. Each aircraft (agent) sends the coordinator information about its intended actions. The coordinator then builds a plan which specifies all of the agents' actions, including the actions that they, or some other node, should take to avoid collisions.

With distributed multi-agent planning, the plan is developed by several agents. This means there may be no one individual which has a global view of the community's activities, hence detecting and resolving undesirable interactions becomes significantly more difficult. Corkill (1979) has developed a distributed hierarchical planner based on NOAH (Sacerdoti, 1977) where agents represent each other using MODEL nodes and plan execution is coordinated by using explicit synchronisation primitives. Rosenschein and Genesereth (1985) use a logic-based approach to study how agents with a common goal, but different local information, can exchange information to converge on identical plans.

Once a plan has been devised, the agents involved are committed to performing the specified actions. If they believed that their acquaintances are unlikely to keep their pledges, then they would not enter the planning phase in the first place because it is such a resource consuming activity. Commitments are, therefore, the foundation of this approach. There is no latitude for deviation from the agreed course of action because it may introduce resource conflicts or other undesirable side-effects which would impair the community's performance. However the situation may change so radically between generation and execution that if the plan was performed the benefit would be negligible or even negative. In this case it is worth entering a

replanning phase to produce a more profitable alternative (Kambhampati and Hendler, 1992; Pollack, 1993). Hence there is a need for conventions to determine when replanning is necessary and whether the existing plan should be reused or whether a fresh plan should be devised.

5 Conclusions

Coordinating the activities of multiple problem solvers is widely regarded as the central problem of DAI research. To be successful a given mechanism requires three facets to be present (Durfee *et al.*, 1989): (i) a structure within which agents can interact in predictable ways; (ii) flexibility so that agents can operate in dynamic environments and can cope with their inherently partial and imprecise view of the community; (iii) appropriate knowledge and reasoning capabilities to intelligently use the structure and flexibility. The Centrality of Commitments and Conventions Hypothesis deals with the first two points. Commitments provide the predictability which agents need to reason about when assessing their role and the role of others in the community. Conventions acknowledge that agents need to respond flexibly, both in their local problem solving and in their interactions towards others, to evolving circumstances. The final feature is not a matter of coordination *per se*; rather it is a factor of the ability of an individual to reason about information and predictions when making decisions about its local problem solving. The aim being to gain the maximum benefit from social interactions while simultaneously contributing to the overall effectiveness of the community.

To argue the case for the Centrality of Commitments and Conventions Hypothesis, the process of coordination was framed in terms of a distributed goal search problem. Using this representation, the fundamental nature of commitments and conventions were demonstrated for a variety of goal-subgoal relationships and goal dependency types. Commitments and joint commitments were presented as fundamental concepts, related to other aspects of an agent's mental state, but not reducible to them. The sole distinguishing characteristic between joint and individual commitments is that the former requires a minimal social convention to ensure the group's status is shared amongst its members; the latter places no restrictions on its associated

conventions.

Recognising commitments and conventions as distinct concepts has the advantage that different types of social system can be represented simply by modifying the convention which is used. The concept of commitment is standard in all applications. Interpretations of higher level semantic notions such as intelligence and rationality can then be encoded using the conventions. In previous work, commitments and conventions were often intertwined and confused with these higher level definitions. The demarcation also highlights the fact that there will be two main sources of agent interaction once the goal tree has been established. One form of interaction will be related to the goal-subgoal relationships and the interagent goal dependencies; the other will be through the specified actions of the conventions. Both of these forms correspond to a different type of social interaction and so provide an aid for structuring the ongoing work related to the types of communication and social processes which are appropriate in different circumstances.

The review of those coordination models which explicitly encode commitments and conventions revealed that the majority of the formalisms concentrate on the notion of commitment, there has been significantly less explicit work on conventions. Three other major coordination mechanisms (organisational structuring, meta-level information exchange and multi-agent planning) were also analysed and their success in producing coherent group behaviour was attributed to the degree of commitment they provide. They represented commitments over varying time horizons (organisations long term, multi-agent plans short term) and of varying levels of detail (from the approximate distribution of tasks present in the organisational structure to the precise details of each agents' actions in multi-agent plans). Again these mechanisms have paid less attention to explicit conventions.

The framework presented here represents an initial step towards a canonical theory of DAI based upon the concepts of commitments and conventions. However much work still needs to be done to draw together the diverse strands into an integrated and complete description of multi-agent behaviour. Some of this work needs to address basic methodological shortcomings:

• The formal theories need to use concepts and languages which can more easily be mapped into computational systems.

• Clear functional and implementation architectures based on the social roles sketched out for commitments and conventions need to be devised.

• Richer multi-agent programming languages, which have a firm theoretical underpinning, need to be developed.

Other work is also needed to clarify the open issues which still surround the commitment and convention framework. Key questions which need to be answered include:

• What mechanisms can be employed to force agents to honour their commitments or to stick to their conventions?

• What are the key domain parameters which determine the appropriate balance between predictability and flexibility to be designed into the conventions?

• How can agents coordinate their behaviour effectively without requiring full mutual belief of each others' commitments and conventions?

• What are the most appropriate mechanisms for obtaining agreements about the conventions to use in a particular situation?

Acknowledgments

This document has benefitted enormously from discussions with the following people: Phil Cohen, Ed Durfee, Les Gasser, Jeff Pople and Mike Wooldridge. I would also like to thank John Fox for his editorial suggestions.

References

Aarnts, R P, Corera, J, Perez, J, Gureghian, D and Jennings, N R, 1991. "Examples of Cooperative

Situations and their Implementation", *Vleermuis Journal of Software Research* **3** (4) pp 74- 81.

Agha, G 1986. *ACTORS: A Model of Concurrent Computation in Distributed Systems*, MIT Press.

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

Audi, R, 1973. "Intending", *The Journal of Philosophy* **70** pp 387-403.

Becker, H S, 1960. "Notes on the Concept of Commitment", *American Journal of Sociology* **66** (1) pp 32-40.

Bond, A H, 1989. "Commitment: Some DAI insights from Symbolic Interactionist Society", *Proc. of 9th Workshop on Distributed Artificial Intelligence*, USA.

Bond, A H and Gasser, L, eds. 1988. *Readings in Distributed Artificial Intelligence*, Morgan Kaufmann.

Bratman, M E, 1984. "Two Faces of Intention" *Philosophical Review* **93** pp 375-405.

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

Bratman, M E, Israel, D J and Pollack, M E, 1988. "Plans and Resource Bounded Practical Reasoning" *Computational Intelligence* **4** pp 349-355.

Burmeister, B, and Sundermeyer, K, 1992. "Cooperative Problem Solving Guided by Intentions and Perception" In E. Werner and Y. Demazeau (eds.) *Decentralised A.I. 3*, Elesevier, pp 77-92.

Cammarata, S, McArthur, D and Steeb, R, 1983. "Strategies of Cooperation in Distributed Problem Solving", *Proc. Int. Joint Conf. on Artificial Intelligence*, Karlsruhe, Germany pp 767-770.

Clearwater, S H, Huberman, B A, and Hogg, T, 1991. "Cooperative Solution of Constraint

Satisfaction Problems" *Science* **254** pp 1181-1183

Chaib-Draa, B, Moulin, B, Mandiau, R and Millot, P, 1992. "Trends in Distributed Artificial Intelligence", *Artificial Intelligence Review* **6** pp 35-66.

Cockburn, D, Varga, L Z and Jennings, N R, 1992. "Cooperating Intelligent Systems for Electricity Distribution", *Proc. Expert Systems 1992* (Applications Track), Cambridge, UK.

Cohen, P R and Levesque, H J, 1990. "Intention is Choice with Commitment" *Artificial Intelligence* **42** pp 213-261.

Cohen, P R, and Levesque, H J, 1991a. "Teamwork", *Nous* **25** (4) (also appears as SRI Technical Note 504, Menlo Park, CA).

Cohen, P R, and Levesque, H J, 1991b, "Confirmations and Joint Action", *Proc. Twelfth Int. Joint Conf. on Artificial Intelligence*, Sydney, Australia.

Conte, R, Miceli, M, and Castelfranchi, C, 1990. "Limits and Levels of Cooperation: Disentangling Various Types of Prosocial Interaction" *Proc. Modelling an Autonomous Agent in a Multi-Agent World*, Saint-Quentin en Yvelines, France.

Corkill, D D, 1979. "Hierarchical Planning in a Distributed Environment", *Proc. Sixth Int. Joint Conf. on Artificial Intelligence*, Cambridge, USA pp 168-175.

Corkill, D D and Lesser, V R, 1983. "The use of meta-level control for coordination in distributed problem solving" *Proc Int. Joint Conf. on Artificial Intelligence*, Karlsruhe, Germany, pp 748-756.

Cox, B J, 1990. "Planning the Software Industrial Revolution" *IEEE Software*, Nov. 90 pp 25-33.

Davidson, D 1980. *Essays on Actions and Events*, Oxford University Press.

Decker, K S, 1987. "Distributed Problem Solving Techniques: A Survey" *IEEE Trans. on Systems Man and Cybernetics* **17** (5) pp 729-740.

Dennett, D C 1987. *The Intentional Stance*, Bradford Books / MIT Press.

Durfee, E H, 1988. *Coordination of Distributed Problem Solvers*, Kluwer Academic Publishers.

Durfee, E H, Lesser, V R, and Corkill, D D, 1987. "Coherent Cooperation among Communicating Problem Solvers" *IEEE Trans. on Computers* **36** pp 1275-1291.

Durfee, E H, Lesser, V R, and Corkill, D D, 1989. "Trends in Cooperative Distributed Problem Solving", *IEEE Trans. on Knowledge and Data Engineering* **1** (1) pp 63-83.

Durfee, E H, and Montgomery, T A, 1991. "Coordination as Distributed Search in a Hierarchical Behaviour Space", *IEEE Trans. on Systems Man and Cybernetics* **21** pp 1363-1378.

Fikes, R E, 1982. "A Commitment-Based Framework for Describing Informal Cooperative Work", *Cognitive Science* **6** pp 331-347.

Galbraith, J 1973. *Designing Complex Organizations*, Addison-Wesley.

Galliers, J R, 1988. "A Strategic Framework for Multi-Agent Cooperative Dialogue" *Proc. European Conf. on Artificial Intelligence*, Munich, Germany, pp 415-420.

Gasser, L, 1991. "Social Conceptions of Knowledge and Action: DAI Foundations and Open System Semantics" *Artificial Intelligence* **47** pp 107-138.

Gasser, L, 1992a. "An Overview of DAI" In: N. M. Avouris & L. Gasser (eds.) *Distributed Artificial Intelligence: Theory and Praxis*, Kluwer Academic Publishers pp 9-30.

Gasser, L, 1992b. "DAI Approaches to Coordination" In: N. M. Avouris & L. Gasser (eds.) *Distributed Artificial Intelligence: Theory and Praxis*, Kluwer Academic Publishers pp 31-51.

Gasser, L, and Huhns, M N, eds. 1989. *Distributed Artificial Intelligence Vol II*, Pitman Publishing

Georgeff, M, 1983. "Communication and Action in Multi-Agent Planning" *Proc. of National Conf.*

on Artificial Intelligence, Washington, DC, pp 125-129.

Gerson, E M, 1976. "On Quality of Life", *American Sociological Review* **41** pp 793-806.

Gilbert, M, 1989. *On Social Facts*, Routledge.

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

Halpern, J Y, 1986. "Reasoning About Knowledge: An Overview", In: J. Y. Halpern (ed.) *Theoretical Aspects of Reasoning About Knowledge* Morgan Kaufmann, pp 1-17.

Halpern, J Y, and Moses, Y O, 1984. "Knowledge and Common Knowledge in a Distributed Environment", *Proc. of the Third ACM Conf. on Principles of Distributed Computing*, pp 50-61.

Harel, D, 1984. "Dynamic Logic", In: D. Gabbay and F. Guenthner (eds.) *Handbook of Philosophical Logic Vol II*, Reidel Publishing, pp 497-604.

Hayes-Roth, F, 1980. "Towards a Framework for Distributed AI", *SIGART Newsletter* pp 51-52.

Hern, L E C, 1988. "On Distributed Artificial Intelligence" *The Knowledge Engineering Review* **3** (1) pp 21-57

Hewitt, C E, and Kornfield, W A, 1980. "Message Passing Semantics", *SIGART Newsletter* pp 48.

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

Huberman, B A and Hogg, T, 1988. "The Behaviour of Computational Ecologies", In: B. A. Huberman (ed.) *The Ecology of Computation*, North Holland, pp 77-115.

Huhns, M N, ed. 1988. *Distributed Artificial Intelligence*, Pitman Publishing.

Huhns, M N, Mukhopadhyay, U, Stephens, L and Bonnell, R, 1988. "DAI for Document

Retrieval" In: M. N. Huhns (ed.) *Distributed Artificial Intelligence*, Pitman Publishing pp 249-284.

Ishida, T., Yokoo, M., and Gasser, L., 1990. "An Organisational Approach to Adaptive Production Systems", *Proc of 8th National Conf. on Artificial Intelligence*, Boston, USA, pp 52-58.

Jennings, N R, 1992. "Towards a Cooperation Knowledge Level for Collaborative Problem Solving", *Proc. 10th European Conf. on Artificial Intelligence*, Vienna, Austria, pp 224-228.

Jennings, N R, 1993. "*Specification and Implementation of a Belief-Desire-Joint Intention for Collaborative Problem Solving*", KEAG Technical Report, Dept. Electronic Engineering, Queen Mary and Westfield College, University of London.

Jennings, N R and Mamdani, E H, 1992. "Using Joint Responsibility to Coordinate Collaborative Problem Solving in Dynamic Environments", *Proc of 10th National Conf. on Artificial Intelligence*, San Jose, USA, pp 269-275.

Jennings, N R, Mamdani, E H, Laresgoiti, I, Perez, J, and Corera, J, 1992. "GRATE: A General Framework for Cooperative Problem Solving" *Journal of Intelligent Systems Engineering* **1** (2) pp 102-114.

Jennings, N R and Wittig, T, 1992. "ARCHON: Theory and Practice", In N. M. Avouris and L. Gasser (eds.) *Distributed Artificial Intelligence: Theory and Praxis*, Kluwer Academic Press pp 179-196

Kambhampati, S and Hendler, J A, 1992. "A Validation Structure Based Theory of Plan Modification and Reuse" *Artificial Intelligence* **55** pp 193-258.

Kinny, D N and Georgeff, M. P., 1991. "Commitment and Effectiveness of Situated Agents", *Proc. Int. Joint Conf. on Artificial Intelligence*, Sydney, Australia, pp 82-88.

Kinny, D, Ljungberg, M, Rao, A, Sonenberg, E, Tidhar, G, and Werner, E, 1992. "Planned Team

Activity" *Pre-Proceedings of the 4th European Workshop on Modelling Autonomous Agents in a Multi-Agent World*, Rome, Italy.

Lenat, D B, 1975. "BEINGS: Knowledge as Interacting Experts" *Proc. Int. Joint Conf. on Artificial Intelligence,* Tblisi, Russia, pp 126-133.

Lenat, D B, and Feigenbaum, E A, 1991. "On the Thresholds of Knowledge" *Artificial Intelligence* **47** pp 185-250.

Lesser, V R, 1991. "A Retrospective View of FA/C Distributed Problem Solving", *IEEE Trans. on Systems Man and Cybernetics* **21** pp 1347-1363.

Lesser, V R and Corkill, D D, 1983. "The Distributed Vehicle Monitoring Testbed: A Tool for Investigating Distributed Problem Solving Networks" *AI Magazine* pp 15-33.

Lesser, V R, and Corkill, D D, 1987. "Distributed Problem Solving", In: S. C. Shapiro (ed.) *Encyclopedia of Artificial Intelligence* John Wiley and Sons pp 245-251.

Levesque, H J, Cohen, P R and Nunes, J H, 1990. "On Acting Together" *Proc. 8th National Conf. on Artificial Intelligence*, Boston, USA, pp 94-99.

Lochbaum, K E, Grosz, B J, and Sidner, C L, 1990. "Models of Plans to Support Communication" *Proc. 8th National Conf. on Artificial Intelligence*, Boston, USA, pp 485-490.

Malone, T W, 1987. "Modelling Coordination in Organizations and Markets" *Management Science* **33** pp 1317-1332.

March, J G and Simon, H A 1958. *Organisations*, Wiley.

McClelland, J L and Rumelhart, D E 1986. *Parallel Distributed Processing*, MIT Press.

McDermott, J, 1990. "Developing Software is Like Talking to Eskimos about Snow" *Proc. 8th National Conf. on Artificial Intelligence*, Boston, USA, pp 1130-1133.

Mead, G H 1934. *Mind, Self and Society*, University of Chicago Press.

Minsky, M 1985. *The Society of Mind*, Simon & Schuster.

Neches, R, Fikes, R, Finin, T, Gruber, T, Patil, R, Senator, T and Swartout, T, 1991. "Enabling Technology for Knowledge Sharing", *AI Magazine* pp 36-56.

Pan J Y C and Tenenbaum, J M, 1991. "An Intelligent Agent Framework for Enterprise Integration" *IEEE Trans. on Systems Man and Cybernetics* **21** pp 1409-1419.

Papazoglou, M P, Laufman, S C and Sellis, T. K., 1992. "An Organisation Framework for Cooperating Intelligent Information Systems" *Journal of Intelligent and Cooperative Information Systems* **1** pp 169-202.

Pollack, M E, 1990. "Plans as Complex Mental Attitudes", In: P.R.Cohen, J.Morgan and M.E.Pollack (eds.) *Intentions in Communication* MIT Press, pp 77-105.

Pollack, M E, 1992. "The Uses of Plans", *Artificial Intelligence* **57** pp 43-68.

Power, R, 1984. "Mutual Intention", *Journal for the Theory of Social Behaviour* **14** pp 85-105.

Rao, A S and Georgeff, M P, 1991. "Modelling Rational Agents within a BDI Architecture", In R. Fikes and E. Sandewall, (eds.), *Int. Conf. on Principles of Knowledge Representation and Reasonin*g, Morgan Kaufmann.

Rao, A S, Georgeff, M P, and Sonenberg, E A, 1992. "Social Plans: A Preliminary Report", In: E. Werner and Y. Demazeau (eds.), *Decentralised A I 3*, North Holland, pp 57-76.

Rosenschein, J S and Genesereth, M R, 1985. "Deals among Rational Agents", *Proc. 9th Int. Joint Conf. on Artificial Intelligence*, Los Angeles, USA, pp 91-99

Sacerdoti, E D, 1977. *A Structure for Plans and Behaviour*, Elsevier.

Searle, J R 1983. *Intentionality: An Essay in the Philosophy of Mind*. Cambridge University Press.

Searle, J R, 1990. “Collective Intentions and Actions”, In: P.R.Cohen, J.Morgan and M.E.Pollack, (eds.), *Intentions in Communication*, MIT Press pp 401-416.

Shina, S G 1991. *Concurrent Engineering and DFM for Electronic Design*, Van Nostrand Reinhold, New York.

Shoham, Y, 1993. “Agent-oriented Programming” *Artificial Intelligence* **60** pp 51-92.

Shoham, Y and Tennenholtz, M, 1992. “On the Synthesis of Useful Social Laws for Artificial Agent Societies”, *Proc of 10th National Conf. on Artificial Intelligence*, San Jose, USA, pp 276-28.

Simon, H A 1957. *Models of Man*, New-York, Wiley.

Singh, M P, 1990. “Group Intentions” *Proc. of 10th International Workshop on Distributed Artificial Intelligence*, MCC Technical Report ACT-AI-355-90.

Singh, M P, 1992. “A Critical Examination of the Cohen-Levesque Theory of Intentions”, *Proc. 10th European Conf. on Artificial Intelligence*, Vienna, Austria, pp 364-368.

Smith, R G and Davis, R. 1981. “Frameworks for Cooperation in Distributed Problem Solving” *IEEE Trans. on Systems Man and Cybernetics* **11** pp 61-70.

Stefik, M, 1986. “The Next Knowledge Medium” *AI Magazine* **7** (1) pp 34-46.

Tuomela, R and Miller, K, 1988. “We Intentions”, *Philosophical Studies* **53**, pp 367-389.

Werner, E, 1989. “Cooperating Agents: A Unified Theory of Communication and Social Structure”, In: L. Gasser and M. N. Huhns (eds.) *Distributed Artificial Intelligence Vol II*, Pitman Publishing, pp 3-36.

Wesson, R, Hayes-Roth, F, Burge, J W, Stasz, C, and Sunshine, C A., 1981. “Network Structures for Distributed Situation Assessment” *IEEE Trans. on Systems Man and Cybernetics* **11** pp 5-23.