MCC Technical Report Number KBNL-086-93

Intentions for Multiagent Systems

Munindar P. Singh

MCC Nonconfidential

Microelectronics and Computer Technology Corporation
Information Systems Division
3500 West Balcones Center Drive
Austin, TX 78759, USA
(512) 338-3431
msingh@mcc.com

Intentions for Multiagent Systems[*]

Munindar P. Singh[†]

Microelectronics and Computer Technology Corporation
Information Systems Division
3500 West Balcones Center Drive
Austin, TX 78759, USA
(512) 338-3431
msingh@mcc.com

Abstract

Multiagent systems are of interest in several subareas of Cognitive Science and Artificial Intelligence, notably, autonomous agents, multiagent planning and action, discourse understanding, and cooperative work. I motivate and present a formal theory of the intentions of a multiagent system that analyzes them in terms of its internal organization, and the intentions of its member agents. This theory treats social structure directly in terms of the interactions among agents and does not attempt to reduce it to psychological concepts. It makes few assumptions about the architecture of agents and about the manner in which they interact. Thus it is intuitively plausible, can be described in a simple formal model, and is applicable to a large variety of multiagent systems. This theory is applied to an extended example, and some interesting variations are outlined.

[*] This is a considerably revised and extended version of [Singh, 1990a].

[†] This work was partially supported by NSF grant # 8945845 to the Center for Cognitive Science, University of Texas, Austin) and by the Microelectronics and Computer Technology Corporation, Austin.

1 Introduction

Several subareas of Cognitive Science and Artificial Intelligence (jointly called AI in this paper) are concerned with multiagent systems, i.e., with intelligent systems composed of groups of intelligent agents who share a part of the world, and who affect one another through their actions. These subareas include autonomous agents, multiagent planning and action, discourse understanding, and cooperative work [Cohen & Levesque, 1988; Demazeau & Müller, 1990; Georgeff, 1987; Grosz & Sidner, 1990; Hewitt, 1988; Konolige, 1982; Singh, 1991c]. It is convenient to refer to multiagent systems as *groups*. In recent work, I have developed a formal theory of the ability of a group (as ascribed objectively to it) that accounts for its internal structure [Singh, 1991a]. Here I plan to extend and modify those ideas and motivate a formal theory of the intentions of a group of agents.

This theory is meant to apply to groups of agents with a wide range of representational and reasoning capabilities: on one extreme these agents may not be able to reason explicitly at all, and on the other be perfectly rational and have unlimited computational power. This theory aims to make as few stipulations as possible about the architecture of individual agents, and the specific ways in which they may interact. In particular, this theory, unlike the theories of Grosz and Sidner [1990] and Cohen and Levesque [1988], does not attempt to reduce social structure to psychological notions like mutual belief. Thus it is both intuitively plausible, uses a fairly simple formal model, and is applicable to a wide variety of multiagent systems.

Problems with the traditional theories, which motivate this paper, are discussed in detail in §3; however, some clarification of my general goals is in order here. I shall take it for granted in the sequel that the importance of folk psychological notions such as belief and intention to AI science and engineering is established [Dennett, 1987; McCarthy, 1979]. Briefly, the reasons for this are the following. (1) Beliefs and intentions provide powerful abstractions with which we, *qua* designers and analyzers, can succinctly describe, understand and explain the behavior of complex systems. (2) Beliefs and intentions make available certain regularities and patterns of action that are independent of the exact physical implementation of the agents. However, for these concepts to be effectively used in both the science and engineering of multiagent systems, they must be given an objective grounding in terms of the architectures that different kinds of systems have and the actions they perform. My aim in this paper is to make the social structure of a group of agents explicit (in a sufficiently abstract model) and to use it, along with some description of the agents themselves, to give a formal model-theoretic definition of the intentions of groups. I shall argue that this approach has advantages over traditional approaches in which social structure is not considered explicitly at all, and is implicitly reduced to further psychological notions.

Briefly, in this paper, the intentions of a group are described in terms of its internal "structure," and the abstract "strategy" it may be said to be following. The quoted terms are described and formalized later in this paper. Roughly, the structure of a multiagent system is what distinguishes, e.g., a commander from a private in an army, and a catcher from a pitcher in a baseball team. The strategy of a group describes what it is doing. This strategy is itself seen as the set of strategies of its members. The structure of a group is captured in terms of the interactions of its members as they follow their respective strategies [Singh, 1991a]. As a consequence, this theory, unlike those of Grosz and Sidner [1990] and Cohen and Levesque [1988], does not require mutual beliefs among the members of a group. This is important because mutual beliefs are impossible to achieve in most realistic scenarios, e.g., where communication delay is not bounded or the communication channels are not reliable [Fischer & Immerman, 1986; Halpern & Moses, 1987].

The theory developed in the paper may be used in the design and analysis of multiagent systems in the following way. It recursively partitions the task of modeling such systems into the tasks of modeling their internal structure and the strategies of their members. The strategies of members are defined simply using a notation related to regular programs [Kozen & Tiurzyn, 1990] and allow us to succinctly describe the relevant aspects of the agents' design (see §5). These strategies yield the intentions of the agents in a simple and direct way. This turns out to be quite important. Usually, modal approaches to intentions or beliefs simply postulate a primitive alternativeness relation that captures the relevant dispositions of the given agents. However, it is not always clarified how such a relation may be implemented in the agent's design. When strategies as defined here are used also as proposed here, this connection is at once rigorous and obvious. A theory such as this one can be used to motivate design rules for multiagent systems, by suggesting what the strategies of the members of a group and its internal structure ought to be for it to have the relevant intentions. The intentions of agents are needed not only to predict or explain their behavior abstractly, but also to give a formal semantics for different kinds of communication among them [Singh, 1991c].

In §2, I present some observations and intuitions about multiagent systems construed abstractly, especially their internal structure, the actions they perform and the intentions they have. In §3, I briefly review the traditional theories of group actions and intentions, and point out some conceptual problems with them relative to the intuitions presented in §2. In §4, I present the major methodological motivations behind this work, a clarification of which is useful in placing this work better in the context of multiagent systems in AI. In §5, I review some relevant parts of a previous paper [Singh, 1991a] extending it as needed for the purposes of this paper. In §6, I present the new theory of group intentions, apply it to an extended example, and outline some interesting variations on the basic definition that account for some senses of group intentions that would be needed for different AI applications. In §7, I describe some of the logical consequences of this theory.

2 Intuitions About Groups

Groups are *structured.* A group's actions and intentions depend greatly on its internal structure, and on the intentions and skills of its members. A group may be said to perform an action even if only one or a few of its members are actually responsible for that action; e.g., we may say that a software group has created two pieces of software, even though no member of the group worked on both, and some worked on neither. In a large corporation, most employees would not even be aware of the identities of most other employees, though they would all contribute to the corporation's actions. An extreme example is an ant colony or bee hive where the member agents may not even be aware of being agents or being members in a group.

Similarly, one can say that a group intends to achieve some objective even though only some members of the group really intend it. E.g., a house building team may be said to intend to put the roof after the walls, while in fact the wall building crew might just move off at that point to a new job (and not care about the roof of the house they were working on). Again, some of the members need not know who the other members are. Often, a group functions successfully as a group only because of the social or sociobiological structure it embodies, not because of explicit reasoning by its members. Examples of such groups are bureaucracies (which mostly process documents routinely), ant colonies and bee hives. The members need not be aware of this structure, and it may change without any explicit reasoning on their part about it, e.g., when some features of the environment change, as might happen when there is a fire under a bee hive.

The members of a group need not have the same intention as the group—they may even be aware of the group's intention but have an incompatible intention. E.g., an army may include some drafted pacifists, or some spies of its enemy—the army's intentions are perhaps not affected much, unless the pacifists and spies hold high ranks (although its chances of success may be reduced). In other cases, explicit dissent is normal; e.g., the U.S. Senate intends whatever legislation more than half of its members do, so several members may not have the intention of the group. Similarly, the United Nations Security Council intends to censure a nation if and only if a majority of its members do, and none of its five permanent members object.

The form of heterogeneity of groups that matters for this paper is the kind that arises due to the members of a group having intentions significantly different than the intentions of the group as a whole. This could happen in groups where the members' intentions are orthogonal or even opposite to the intentions of the group as a whole. E.g., a free market economy may be ascribed the intention of moving goods efficiently from producers to consumers, while all its participants intend as individuals is to make a profit.

Two observations, made in [Singh, 1991a], are also in order here.

1. **External Monolithicity:**

   A group (e.g., a sports team) may be considered as a single unstructured monolithic agent from without; i.e., groups are "Hobbesian corporate persons," in Hamblin's term [1987, pp. 60, 240]. The intention of a group of agents is the same type of entity as the intention of individual agents: the only difference is one of extent—one would typically expect a cooperative group to have an intention that no proper subgroups of it could succeed with. One consequence of this is that groups, e.g., business corporations, may recursively contain other groups.

   Theoretically, the important consequence of this is that single agent systems ought to turn out merely to be special cases of multiagent systems. This is not a question of just theoretical elegance, however. It is crucial that we be able to reason hierarchically about groups of agents, so that what is taken to be a single agent at one level of abstraction may be considered an internally structured system at another level of abstraction. We would be in a pretty awkward position if merely going up or down a level of abstraction in describing a multiagent system changed the semantics of the intentions ascribed to the system, e.g., if we could say of the U.S. Senate that it intends to raise taxes when seen as a monolithic group, but not when seen as having 100 members, of whom only 90 intend to raise taxes.

2. **Internal Heterogeneity:**

   Even though groups, e.g., teams and corporations, may seem monolithic from without, they are not homogeneous. Typically, the members of a group do not just have different abilities, but also make differing contributions to the intentions of the group, depending on the manner in which they participate in it; e.g., the U.N. Security Council is structured so that its five permanent members are distinguished from its temporary members. Groups may have further structure due to nesting; e.g., large corporations and armies are groups whose constituents (departments and divisions, respectively) are also groups.

3 The Traditional Theories

The most well-known AI theories of group intentions are the ones of Grosz and Sidner [1990], and Cohen and Levesque [1988]. These theories seem to have been developed for the domain of discourse understanding, although they have been proposed as theories of the intentions of multiagent systems at large. In this paper, I examine them as the latter.

Grosz and Sidner's theory is developed in the context of "shared plans," which are plans that a group of agents may have in order to do a particular action. The agents are required to have a mutual belief that they have the given shared plan, and a mutual belief that each agent intends to do his part of it and by doing his part to satisfy the shared plan. Cohen and Levesque's theory is quite complicated and is only partially described here. They define a joint persistent goal of a group of agents as a goal that (1) the members of the group mutually believe is a goal of each of them, and (2) which they will not give up till they mutually believe it has been satisfied or mutually believe has become unsatisfiable. A joint intention for them is simply a particular kind of joint persistent goal, namely, one that is for the following proposition: the agents achieve the intended condition and, until it is achieved, they mutually believe that they are about to achieve it.

For our purposes, the most important feature of these theories is that they both require that for a group to have an intention, its members should have a *mutual belief* that they all have that intention (or "goal"). Roughly, a mutual belief of a set of agents in something means that each of them believes it, and each of them believes that each of them believes it, and so on, *ad infinitum.* Cohen and Levesque additionally require that a member who drops his intention must inform others that he has done so.

Not only is the mutual belief requirement computationally demanding (so that agents may reason about others to arbitrary nesting of beliefs), it also requires a lot of communication among the members (for the mutual beliefs to be established). In general, mutual beliefs are impossible to establish using asynchronous communication [Fischer & Immerman, 1986; Halpern & Moses, 1987]. In practice, they can be established only if certain conventions are stipulated.

Most importantly, however, the mutual belief requirement makes these analyses applicable to a concept different, and more complex, than simple intentions. The member agents of a group are required to be aware of each other's intentions, and beliefs about intentions, and to be aware that each of them is so aware, and so on. Thus the concept described seems to be that of, what one might call, *"perfectly introspective intention,"* rather than intention *simpliciter*. Indeed, if we consider a group consisting of exactly one member, we would expect that the intentions and beliefs of the group would be the same as that of its only member. However, under the traditional accounts cited above, the intention of the group corresponds to the conjunction of beliefs, to arbitrary nesting, that the member has that intention—clearly, a very different concept.

The traditional theories are also unable to account for the internal structure of groups. They assume that all agents are equally capable (in terms of the "basic" actions they can do—they are usually allowed differing knowledge of facts and differences in capabilities that occur solely as a consequence of differences in knowledge of facts). Moreover, traditional theories require the members of a group to be perfectly cooperative in that they use some notion of a "mutual goal"—a goal that the agents share along with a mutual belief that they share that goal. By requiring mutual beliefs these theories, in effect, require that groups be homogeneous—this is because a mutual belief is symmetric with respect to all the agents who have it. Note that the kind of homogeneity these theories require is homogeneity with respect to the intentions of the group, as explained in §2. As a consequence, these definitions do not apply even if (1) one member of the group dissents, or (2) one member of the group for some reason does not believe he has the same intention as the group, or (3) one member of the group doubts the "loyalty" of one of his fellow members, or (4) one member is not convinced that condition (3) above does not hold for some member to some arbitrary level of nesting, or any of such conditions holds. Such groups would be too unstable in practice—that real-life groups are not always so unstable suggests that these theories are somewhat restrictive.

Traditional theories identify a group with the set of its members. They are also committed to a plan-based view of intelligence, a view that has come under much criticism recently [Agre & Chapman, 1987]. The arguments of Agre and Chapman include the following: (1) Symbolic plans are expensive to represent and interpret—many intelligent agents would in fact not be able to use them because of their limitations. (2) Symbolic plans must of necessity be incomplete. Though these arguments are made in an informal framework they are pertinent here, since we would not want to limit our theory of intentions to agents of unlimited computational power.

But more to the point here is a criticism that, as far as I know, is novel to this paper. The traditional view of intentions as symbolically represented plans requires that the agents' goals or plans be explicitly represented in them. Whatever its merits for the single agent case, this approach leaves us with the following problem for groups: *Where do the explicit representations of the plans of a group lie?* There is no single "brain" where these plans may be located. In a way, the idea of the papers of Grosz and Sidner, and Cohen and Levesque to introduce mutual beliefs may be seen as an attempt to answer this very question: For them, a group intends something only if its members mutually believe that each of them has that intention. As already shown in this section, this leads to several problems. Intuitively, it seems that the reason why the mutual belief option has seemed natural to other researchers is that they do not assign sufficient importance to the social structure of groups, but rather focus on only the psychological states of their members. Mutual beliefs are then the obvious choice.

I shall argue that the traditional theories apply to what is merely a special case of the general concept of group intentions. Mutual belief is a requirement that, if it ever applies,

applies only to some particular kinds of multiagent systems. Similarly, the requirement of Cohen and Levesque's theory that agents inform others when they drop their intentions is only a special convention that might make sense some of the time, but not always. One possible explanation for this clash of intuitions is that arguments applicable to discourse situations do not easily extend to all kinds of group action. Discourses require actions whose main effects are *within* the given group, on the mental states of the participants; in other situations, the effects of actions *outside* the group, and in the real world, matter the most. However, the traditional theories cited are claimed to be theories of intentions of groups at large. Therefore, in brief, the objections to these theories are that they

1. Define the concept of "introspective" intention instead of simple intention.
2. Ignore group structure, and cannot account for groups that are nested.
3. Try to reduce social phenomena to psychological concepts like mutual belief.
4. Require special conventions to hold (for mutual beliefs to be established).
5. Require excessive communication (also for mutual beliefs to be established).
6. See intentions as exclusively symbolically represented.
7. Require agents to be computationally and representationally powerful.

As will become clear shortly, these problems are avoided by the theory presented here.

4 Methodological Considerations

I ought to clarify the meanings of two terms I use often in the sequel: "internal" and "external." By the former I mean any theory or approach that takes the viewpoint of the agent being analyzed or specified. By the latter I mean any theory or approach that takes the viewpoint of an external observer, and tries to characterize a system in terms of its behavior as objectively observed. Internal approaches speak of the actual structures that exist within a system. External approaches need to ascribe internal structures also, but only to the extent that they are forced to do so by objective observations. In its basic form, the theory developed here is meant to be used *about* agents rather than *by* them. My goal in the research reported here is to provide a rigorous foundation for the ascription of mental states, especially intentions, to different kinds of agents.

The approach of this paper is model-theoretic in the standard sense of logic, as applied in AI or elsewhere. Statements of fact (including statements of what a given agent intends) are evaluated with respect to a formal "model" in which different possible states of the world are involved (e.g., see [Chellas, 1980, pp. 34–35]). Whether the statement "it is raining" is true in the model or not depends only on the state of the world relative to which this statement is evaluated, *not* the beliefs of any agent. Similarly, whether an agent intends something is to be differentiated from the question of whether he (or someone else) believes that he intends something. This point is obvious and standard, but can cause grave misconceptions if not kept in mind.

As I explained above, I address the problem of externally *ascribing* intentions to agents. Firstly, the ascription is external in that it is not produced by the agents themselves—this means that there must be some principled objective grounds for making these ascriptions. Secondly, if that helps, these are just ascriptions; i.e., one could, in principle, take the stance that these agents do not *really* have any intentions, even though they might be attributed some. However, I do not recommend the latter stance, but not for any philosophical reasons—my reasons are entirely technical or pragmatic. I discuss these points in turn below.

The "principled" grounds for ascribing intentions that I alluded to above have something to do with how we might *validate* claims of ascription. The idea is that one would like to exclude gratuitous ascriptions of intentions as far as possible, and work under the constraint that the actions of agents be considered explicitly; in other words, agents' intentions should be judged by their deeds rather than anything else. In an internal stance, it is possible for an agent to have an intention that is never acted upon; in the stance taken here that would be impossible. This approach, while it gives importance to behavior, is not behavioristic. This is because (1) it does not just consider the behavior in the given situation, but in all possible ones and (2) as is described in detail in §5.3, it considers the internal architecture of both individual agents and groups.

The idea that claims of intentions and beliefs ought somehow to be validated by the agent's actions is not novel to this paper. It is a standard claim of *pragmatism*, the philosophy behind the formal logic and semantics that is based on "possible worlds" models [Stalnaker, 1984, pp. 15–19]. Such models are used in a number of formal theories, e.g., those of [Fischer & Ladner, 1979], [Halpern & Moses, 1987] and this paper (see [Chellas, 1980] for a textbook level introduction). This idea is also given importance in other "naturalist" frameworks, e.g., the one of Barwise and Perry [1983].

From the technical logical point of view, what distinguishes this approach from an internal approach is that all the evidence we may have for ascribing (to a given agent) an intention toward a particular condition can also be used as evidence for ascribing an intention toward the logical, i.e., logically necessary, consequences of that condition. In an internal approach, of course, one would not desire such a conclusion—the agent who intends a condition might not be aware of its logical consequences, or be aware of them, but not intend them for other reasons. Note that all approaches based on standard modal logic permit this inference (or a related one involving logical equivalence, rather than consequence). The reason this inference is acceptable from an external viewpoint is that we are committed to considering the actions of the agent in terms of their effects in the model, and there is no obvious principled way of excluding the necessary consequences of an agent's intentions.

Thus the theory proposed here applies in quite the same way as McCarthy's ascriptions of beliefs and intentions to a thermostat [1979, pp. 11–12]. Rosenschein [1985] also treats the knowledge of an agent in terms of the objective correlations between the agent's state and its environment, and explicitly allows that the agent know all the consequences of its knowledge. This is thus the same sense of "know" that Barwise and Perry use in "...how the praying mantis, which grows by startling spurts, still *knows* precisely how far it must reach to grab a prey" [1983, p. 11, emphasis added]. Just as for intentions, this would be disastrous from an internal point of view, but from an external perspective, it seems intuitively quite acceptable—the agent's actions that are compatible with his knowing a fact are also compatible with his knowing its logical consequences.

The other point I mentioned earlier in this section was about whether agents really have intentions. An alternative view, intuitively quite plausible when one is talking of artifacts created by humans, is that agents (e.g., thermostats) do not have any intentions of their own, but merely reflect their designer's "intent." But there are several problems with taking such a view seriously in ones theorizing:

- There are multiagent systems for which one can identify no unique designer—e.g., the British Parliament, or ones favorite society.
- The designer's intent would involve *types* of situations (e.g., "in conditions of serious foreign threat, the Parliament should intend to suspend the freedom of speech of the people"), while the intentions of the agents, as we need that concept, involve *tokens* or specific conditions (e.g., "on October 1, 1940, the Parliament intends to suspend the freedom of speech of the people").
- The designer usually is not around to supervise the functioning of the given system. If a system can be considered autonomous, it can, and must, be ascribed intentions of its own. The situations an agent faces can include those that were not anticipated by the designer. Additionally, if the designer's intent really mattered, no system would ever perform incorrectly—as remarked above, we must evaluate agents by their actions and validate claims about their intentions accordingly.

Further desiderata for a theory of group intentions are the following. A theory of intentions should not be committed to a plan-based architecture of intelligent agency, since intelligence is not solely a matter of explicitly representing and interpreting symbolic structures, or at least not necessarily so [Agre & Chapman, 1987]; it should, however, be compatible with a plan-based architecture. A useful theory would accommodate the idea of situated action and also consider the interactions among a group's members as they emerge from collective action. The attribution of intentions to a group of agents depends not only on their psychological state, but also on their habits and skills, as well as the social interactions they have among themselves. These are aspects that cannot be easily reduced to psychological concepts. And as I hope to show, if one gives up the requirement to reduce all aspects of behavior to purely psychological concepts, one can obtain a theory that is perspicuous, and yet captures many of our pre-theoretic intuitions about intentions. Such a theory would also give an account of the psychological phenomena that relates the latter directly to the behavior and the architecture of groups, rather than requiring both of them to be independently attributed in an *ad hoc* manner.

The first sense of intentions considered in this paper (in §6) is a purely external notion, and is the core notion introduced by this paper. It may be applied to a variety of systems, though it would be the most useful for systems for which a simpler physical description is not available. The other senses of intentions considered in this paper provide a more realistic picture of ascription but, in turn, are more complex.

5 The Framework

I now turn to a description of the formal model, and the auxiliary definitions on which the theory of this paper will be based. While this section is essential for a thorough understanding of the definitions to follow, it may safely be skimmed over on a first reading.

5.1 The Formal Model

The formal model here is related to the ones in [Singh, 1990b; Singh, 1991a]; an important difference is that the interpretation assigns strategies (that are 'had'), conditions (that are 'believed') and pairs of strategies and conditions (in which the strategies are 'intended-for' the conditions) to agents. Some new predicates used in this paper are also defined.

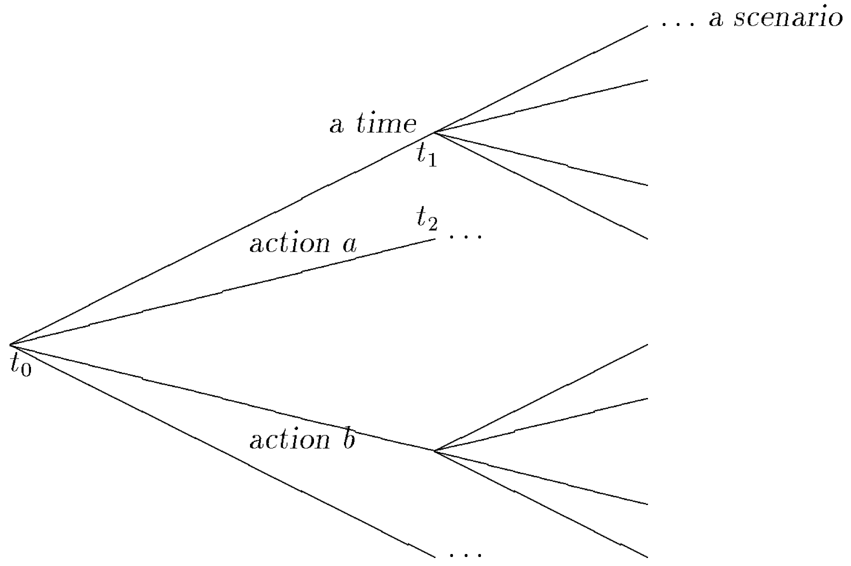


Figure 1: A World with a Branching History

The formal model is based on possible worlds. As diagramed in Figure 1, each possible *world* has a branching *history* of *times*. At each time, environmental processes, and agents' actions occur. Agents influence the future by acting, but the outcome also depends on other events; e.g., in Figure 1, the agent can constrain the future to some extent by choosing to do action a or action b, but what exactly transpires, e.g., whether t_1 or t_2 becomes the case, cannot be controlled by him. I allow the actions an agent can do to be different at different times.

Let $M = \langle F, N \rangle$ be an intensional model, where $F = \langle \mathbf{W}, \mathbf{T}, <, \mathbf{I}, \mathbf{A}, \mathbf{U} \rangle$ is a frame, and $N = \langle \mathbf{f}, [\![\,]\!], \mathbf{B}, \mathbf{C}^{\mathbf{Y}}, \mathbf{C}^{\mathbf{B}}, \mathbf{C}^{\mathbf{MB}}, \mathbf{C}^{\mathbf{I}} \rangle$ an interpretation. The interpretation is made so complex here only to accommodate several different senses of intentions and to facilitate comparisons with other theories—not all the $\mathbf{C}$'s are needed at once.

In definition of the frame, F, $\mathbf{W}$ is a set of possible worlds, and $\mathbf{T}$ is a set of possible times. $<$ is partial order on $\mathbf{T}$, and represents the temporal order of the times in it. $\mathbf{I}$ is the class of domains of the various possible worlds. $\mathbf{A}$, the class of individual agents in different possible worlds, is a subclass of $\mathbf{I}$. $\mathbf{A}$ is especially important and interesting in this paper since group agents are considered as first class agents in this theory. I shall describe the novel parts of the model after further intuitive motivation in §5.4. The closure of $\mathbf{A}$ under group formation is the class of all agents, $\mathbf{A}^*$. $\mathbf{U}$ is the simply class of basic actions the individual agents, i.e., the members of $\mathbf{A}$, can do.

In the definition of the interpretation, N, $\mathbf{f}$ is a function that at each world and time assigns individuals to individual variables. This is extended in §5.4 to assign group individuals to group variables. $[\![\,]\!]$ assigns intensions to predicates and actions. For an n-ary predicate symbol, ψ, $[\![\psi]\!]$ is a function from $\mathbf{I}^n$ to $\mathbf{W} \times \mathbf{T}$; i.e., the intension of a predicate tells us what worlds and times it is true in when applied to each possible tuple of arguments. The case of actions is described below. $\mathbf{B}$ is the class of functions assigning basic actions (subsets of $\mathbf{U}$) to the agents at different worlds and times. $\mathbf{C^Y}$ is the class of functions assigning strategies to the agents at different worlds and times. These are the strategies the agents 'have' (see §6). Further requirements, such as the persistence of agents with their strategies, may be added if necessary. $\mathbf{C^B}$ is a class of functions assigning conditions or formulae to the agents at different worlds and times. These conditions are *believed* by the agents. $\mathbf{C^{MB}}$ is a class of functions assigning conditions to sets of agents at different worlds and times. These conditions are *mutually believed* by the agents in the given set. For the interpretation to be coherent, $p \in \mathbf{C^{MB}}(x_1, \ldots, x_n) \leftrightarrow (\forall i : (p \in \mathbf{C^B}(x_i)) \wedge (\text{believes}(x_i, p) \in \mathbf{C^{MB}}(x_1, \ldots, x_n)))$. The predicate 'believes' has the obvious meaning and is formally defined in §6.2. $\mathbf{C^I}$ is the class of functions assigning pairs of strategies and conditions to the agents at different worlds and times. Intuitively, each agent 'intends-for' the assigned strategy to be for the assigned condition (see §6). This makes sense only if the strategy in this pair is the same as the one assigned by $\mathbf{C^Y}$, so it assumed that this requirement is enforced by the interpretation. Each world $w \in \mathbf{W}$ has exactly one history, constructed from the times in $\mathbf{T}$. Histories are sets of times, partially ordered by $<$. As diagramed in Figure 1, they branch into the future but are linear in the past. The times in any given history occur only in that history.

A *scenario* at a world and time is any maximal set of times containing the given time, and all times that are in a particular future of it; i.e., a scenario is any single branch of the history of the world which begins at the given time, and contains all times in some linear subrelation of $<$. Define a *skeletal scenario* as an eternal linear sequence of times in the future of the given time; i.e., SS at w, t is any sequence: $\langle t = t_0, t_1, \ldots \rangle$, where $(\forall i : i \geq 0 \rightarrow t_i < t_{i+1})$ (linearity) and $(\forall i, t' : t' > t_i \rightarrow (\exists j : t' \not> t_j))$ (eternity). Now, a scenario, S, for w, t may be defined as the "linear closure" of some skeletal scenario at w, t. Formally, S, relative to some SS, is the minimal set which satisfies the following conditions:

- *Eternity*: $SS \subseteq S$
- *Linear Closure*: $(\forall t'' : t'' \in S \rightarrow (\forall t' : t_0 < t' < t'' \rightarrow t' \in S))$

This definition applies to arbitrary histories, not just discrete ones. $\mathbf{S}_{w,t}$ is the class of all scenarios at world w and time t. $\langle S, t, t' \rangle$ is a subscenario of S from t to t', inclusive.

Basic actions may have different durations in different scenarios, even those that begin at the same world and time. The intension of a predicate is, for each world and time, the set of the relations that model it. The intension of an action is, for each agent x, the set of subscenarios in the model in which an instance of it is done (from start to finish) by x; e.g., $\langle S, t, t' \rangle \in [\![a]\!]_x$ means that agent x does action a in the subscenario of S from time t to t'. An agent (or subgroup) could do several actions at once; since the $[\![\,]\!]$ is a part of the model, the actions that are done simultaneously are automatically mutually compatible. I assume that $[\![\,]\!]$ respects $\mathbf{B}$; i.e., $a \in \mathbf{B}_{w,t}(x)$. The following coherence conditions on models must be imposed for them to be intuitively reasonable [Singh, 1991b]: (1) at any time, an action can be done in at most one way on any given scenario; (2) subscenarios are uniquely identified by the times over which they stretch; i.e., the scenario used to refer to them is not important; (3) there is always a future time available in the model; and (4) something must be *done* by each agent along each scenario in the model, even if it is a dummy action. Although, I will not pursue this idea here, restrictions on $[\![\,]\!]$ can be used to express the limitations of agents, and the ways in which their actions may interfere with those of others; e.g., we might require that x cannot pick three glasses at once, or at most one person can enter an elevator at a time, and so on. Habits of agents, e.g., that x always brakes before turning, can be similarly modeled.

It is useful for the definitions that follow to extend the definition of intension of an action in the following ways. Let $G = \{x_1, \ldots, x_n\}$ be a group of n agents (more conditions will be added to this definition later). Let '$\mathrm{seq}_i = \langle a_0, \ldots, a_{m-1} \rangle$' be a sequence of actions of member x_i. Then $[\![\mathrm{seq}_i]\!] = \{\langle S, t, t' \rangle | (\exists t_0 \leq \ldots \leq t_m : t = t_0 \wedge t' = t_m \wedge (\forall j : j \in [1 \ldots m] \rightarrow \langle S, t_{j-1}, t_j \rangle \in [\![a_{j-1}]\!]_{x_i}))\}$, i.e., the set of subscenarios over which it is done. A sequence for a group is a set of member sequences executed in parallel; formally, $\mathrm{seq} = \langle \mathrm{seq}_1 \parallel \ldots \parallel \mathrm{seq}_n \rangle$. The intension of 'seq' is the set of all subscenarios in which the member sequences are executed starting together (and ending in any order). $[\![\mathrm{seq}]\!] = \{\langle S, t, t' \rangle | (\exists e_0, \ldots e_{m-1} : \{e_0, \ldots e_{m-1}\} = \{0, \ldots, m-1\} \wedge t_{e_0} \leq \ldots \leq t_{e_{m-1}} \wedge (\forall j : j \in [1 \ldots m] \rightarrow \langle S, t, t_{e_j} \rangle \in [\![\mathrm{seq}_j]\!]))\}$. $\langle S, t, t' \rangle \in [\![\mathrm{seq}]\!]$ means that 'seq' starts at t and ends at t'; $[\![\mathrm{seq}]\!] = \emptyset$ means that a subset of the member sequences is inconsistent. Sets of simultaneous actions can be treated as sets of sequences of length unity.

5.2 Formal Language and Semantics

The formal language of this paper, $\mathcal{L}$, is the predicate calculus, augmented with some special predicates including 'intends(agent, formula)' and some temporal operators: A, E, U, F, G and P. $\mathsf{A}p$ means that p holds on all scenarios beginning at the given situation. $\mathsf{E}p$ means that p holds on some scenario beginning at the given situation and abbreviates $\neg\mathsf{A}\neg p$. $p\mathsf{U}q$ means that q holds eventually on the given scenario and p holds until then; $\mathsf{F}p$ means p holds eventually on the given scenario and abbreviates $\mathsf{true}\mathsf{U}p$; $\mathsf{G}p$ means p holds always in the future on the given scenario and abbreviates $\neg\mathsf{F}\neg p$. $\mathsf{P}p$ means that p holds sometimes in the past. For simplicity, both "past" and "future" are taken to include the present time.

The semantics is given relative to intensional models: it is standard for the predicate calculus and the temporal operators; the special predicate 'intends' is considered later in the paper. Some auxiliary predicates, e.g., 'believes' and 'has,' are defined where needed. It is assumed throughout that operators for quoting and dequoting can be inserted where necessary. This simplifies the notation considerably. Using these operators, we can convert between syntactic and semantic objects with ease—this idea is well-known and simple, though it is notationally cumbersome [McArthur, 1988, p. 229].

The novel feature of $\mathcal{L}$ is that it allows agents to be defined out of predefined agents. Recall that the class of agents in the model, $\mathbf{A}^*$, is closed under group formation. For the base case, agents are ordinary individual symbols; groups are then composed out of agents (and are themselves agents). This idea will become clearer in §5.4, where it is formalized. $\mathcal{L}$ is formally defined as the minimal set closed under the following rules. Here Φ is a set of predicate symbols and Ξ is a set of variables.

1. $\psi(x_1, \ldots, x_n) \in \mathcal{L}$, where $\psi \in \Phi$ is an n-ary predicate symbol and $x_1, \ldots, x_n \in \Xi$ are variables.
2. $p, q \in \mathcal{L}$ implies $p \wedge q \in \mathcal{L}$.
3. $p \in \mathcal{L}$ implies $\neg p \in \mathcal{L}$.
4. $p, q \in \mathcal{L}$ implies $p \mathsf{U} q \in \mathcal{L}$.
5. $p \in \mathcal{L}$ and $x \in \Xi$ implies $(\exists x : p) \in \mathcal{L}$.
6. $p \in \mathcal{L}$ implies $\mathsf{A}p \in \mathcal{L}$, where A is a path-quantifier.
7. $p \in \mathcal{L}$ implies $\mathsf{P}p \in \mathcal{L}$, where P is the past operator.

A denotes "in *all* scenarios at the present time." A useful definition is E, which denotes "in *some* scenario at the present time"—i.e., $\mathsf{E}p \equiv \neg\mathsf{A}\neg p$. $p\mathsf{U}q$ means that q sometimes on the future on the given scenario and p holds from now to then. $\mathsf{F}p$ denotes "p holds sometimes in the future on this scenario" and abbreviates "$\mathsf{true}\mathsf{U}p$." $\mathsf{G}p$ denotes "p always holds in the future on this scenario" and abbreviates "$\neg\mathsf{F}\neg p$." $\mathsf{P}p$ denotes "p held at some point in the past." Implication $(p \rightarrow q)$ and disjunctions of formulae $(p \vee q)$ are defined as the usual abbreviations.

The semantics of formulae is given relative to a model, as defined above, and a world and time in it. $M \models_{w,t} p$ expresses "M satisfies p at w, t." $M \models_{S,t} p$ expresses "M satisfies p at time t on scenario S," and is needed for some formulae as defined in §5.2. p is *satisfiable* iff for some M, w and t, $M \models_{w,t} p$. p is *valid* in M iff it is satisfiable at all worlds and times in M. Define $\mathbf{f}^x_{\mathsf{j}}$ as an assignment function that is a variant of $\mathbf{f}$ for variable x; i.e., $\mathbf{f}^x_{\mathsf{j}}(x) = \mathsf{j}$ and $(\forall y \neq x : \mathbf{f}^x_{\mathsf{j}}(y) = \mathbf{f}(y)$. Let M^x_{j} be the model that results by substituting $\mathbf{f}^x_{\mathsf{j}}$ for $\mathbf{f}$ in model M. The satisfaction conditions for the temporal operators too are adapted from those in [Emerson, 1990]. Formally, we have the following definitions:

1. $M \models_{w,t} \psi(x_1, \ldots, x_n)$ iff $\langle w, t \rangle \in [\![\psi]\!](\mathbf{f}_{w,t}(x_1), \ldots, \mathbf{f}_{w,t}(x_n))$
2. $M \models_{w,t} p \wedge q$ iff $M \models_{w,t} p \wedge M \models_{w,t} q$
3. $M \models_{w,t} \neg p$ iff $M \not\models_{w,t} p$
4. $M \models_{w,t} \mathsf{A}p$ iff $(\forall S : S \in \mathbf{S}_{w,t} \rightarrow M \models_{S,t} p)$

5. $M \models_{S,t} p\mathsf{U}q$ iff $(\exists t' : M \models_{S,t'} q \wedge (\forall t'' : t \leq t'' \leq t' \rightarrow M \models_{S,t''} p))$

   $p\mathsf{U}q$ is satisfied at time t on scenario S iff there is a time such that q holds at it, and for all times between now and then, p holds at them.

6. $M \models_{w,t} \mathsf{P}p$ iff $(\exists t' : t' < t \wedge M \models_{w,t'} p)$

7. $M \models_{S,t} p$ iff $M \models_{w,t} p$, if p is not of the form $q\mathsf{U}r$, and w is the (unique) world such that $S \in \mathbf{S}_{w,t}$

8. $M \models_{w,t} (\exists x : p)$ iff there is a $\mathbf{j} \in \mathbf{I}_{w,t}$ such that $M^x_{\mathbf{j}} \models_{w,t} p$

5.3 Strategies

I define "strategies" as the abstract specifications of the behavior of an agent or a group. The formal definition of strategies used here is derived from regular programs in Dynamic Logic—a standard notation for describing programs and computations in theoretical Computer Science [Fischer & Ladner, 1979]. The idea of using strategies such as these for describing intelligent agents can be traced back to McCarthy and Hayes [1969]. However, besides the extension to multiagent systems, strategies in this paper are taken merely to characterize an agent's behavior, possibly in quite coarse terms; i.e., there is no commitment here to strategies being implemented as symbolic structures or as programs—they could just be the compact descriptions of a particular architecture. While the issue of the form that strategies must take is of great importance from the point of view of the implementor, from a logical point of view it is quite immaterial.

Let Y be a strategy of an individual agent x (as assigned by the model); 'current(Y)' the part of Y now up for execution; and 'rest(Y)' the part of Y remaining after 'current(Y)' has been done. These definitions have been borrowed from previous papers [Singh, 1990b; Singh, 1991b]. I now define a strategy, Y, recursively as follows.

0. skip: the empty strategy

1. do(A): a condition to be achieved

2. wait(A): a condition to be awaited, (for synchronization with other events)

3. $Y_1; Y_2$: a sequence of strategies

4. if A then Y_1 else Y_2: a conditional strategy

5. while A do Y_1: a conditional loop

The 'current' part of a strategy depends on the current situation. For cases (0), (1) and (2), 'current(Y)' is 'Y' itself; for (3), it is 'current(Y_1)'; for (4), it is 'current(Y_1)' if A holds in the current situation, and 'current(Y_2)' otherwise; for (5) it is 'current(Y_1),' if A holds (in the current situation), and 'skip' otherwise. The 'rest' of a strategy is what is left after the current part is performed. For cases (0), (1) and (2), 'rest(Y)' is 'skip'; for (3), it is 'rest(Y_1); Y_2'; for (4), it is 'rest(Y_1),' if A holds and 'rest(Y_2)' otherwise; for (5), it is 'rest(Y_1); while A do Y_1,' if A holds, and 'skip' otherwise. It can easily be seen that relative to the standard semantics for the constructs introduced above (e.g., see [Kozen & Tiurzyn, 1990]), 'Y' is equivalent to 'current(Y); rest(Y).' Another obvious consequence of this is that 'current(Y)' is always of the form 'skip' or 'do(A)' or 'wait(A).'

The syntax of strategies as proposed here is meant to be sufficiently powerful to express the actions or architecture of agents who are quite complex. The strategies of agents may be taken to signify their hardware description or the programs they execute. Each individual agent (the most basic kind of agent that I consider) is assigned a strategy in the model and is related to the actions it does in fact. Thus an agent may be assigned only one strategy at a given point in a model.

5.4 Group Structure

Strategies were defined above for single agents. They are also be used in this paper to abstractly characterize the behavior of groups. A group strategy is defined simply as the set of strategies of its members. This is in line with the view that individual agents are initially assigned strategies at different points in the model and that these strategies describe all the relevant aspects of the agents' behavior to sufficient detail. An agent acts in fact as his strategy says he does: he does not act differently for himself than he does for any of the groups he belongs to.

While the strategy of a group is taken as the set of the strategies of its members, the intentions of a group are also constrained by its structure. The structure of a group is defined by the interactions among its members as they perform their actions. The interactions among the members of a group can be seen as objectively determining their respective "roles" in the group. Two kinds of interactions among the members of a group may be distinguished. While they are treated similarly in the formalism, they may have differing imports for the designer or analyzer of a multiagent system.

1. **Strategic Interactions:**

   I define *strategic interactions* as the abstract interactions among the members of groups that are described at the level of their strategies. An important subclass of such interactions involves illocutionary acts between agents, e.g., assertions, commands and promises [Austin, 1962]. Examples of this subclass are the following: (1) in a football team, the receivers run the patterns that the quarterback asks them to (i.e., the receivers obey the quarterback's *commands*); and (2) in the contract net of [Davis & Smith, 1983], one can say the contractors *promise* to do the given task for a certain price by bidding on it. Another subclass of strategic interactions involves the establishment of various conditions in the world by some members' strategies that other members' strategies rely on. Temporal ordering constraints arise in many domains; e.g., (1) in some football plays, some players may be supposed to clear the path for their teammates (i.e., achieve a condition that makes it possible for the other player's strategy to be executed successfully); (2) in football again, whenever a defensive player identifies a move, he would act to block it.

   Strategic interactions may be listed simply by stating the relevant substrategies and the constraints on them. It is interesting to note that the semantics of the illocutionary acts between the members of a group would require an account of their intentions as agents in their own right.

2. **Reactive Interactions:**

   I define *reactive interactions* as the interactions among the members of a group that are not defined in terms of their strategies, but rather just arise as they act. These interactions may be implicit in the way in which a particular group acts and may thus determine the joint "habits" of the group, as it were. E.g., a football player may simply obstruct an opposing player trying to tackle his teammate, even though his strategy is to run forward himself; good football players react to their opponents' moves by pushing as hard as possible without running into their teammates or committing fouls.

   Reactive interactions are crucial to the group's performing its strategy successfully. They may be listed by stating the order of certain basic actions that may be done by the involved agents, even though those actions would not be explicit in their strategies.

This motivates the following definition for a group. Let a group, $G = \langle\langle x_1, \ldots, x_n\rangle, I_S, I_R\rangle$, where the x_i are its *members* (notated, $x_i \in G^m$); and I_S and I_R are, respectively, the sets of *restrictions* on the strategic and reactive interactions among the x_i. Groups are represented in the obvious manner in the syntax using lists of elements, with the constraint that the denotation of a group, as given by the function $\mathbf{f}$ in the interpretation, does not depend on the order in which the elements are enumerated.

The group strategy, Y, is an ordered set of member strategies, $\langle Y_1, \ldots, Y_n\rangle$. The restrictions in $I_S \cup I_R$ must be 'met' as each x_i performs his strategy, Y_i. This ensures that the actions being done are being done by the agents as members of the given group, since each agent plays the appropriate role. The actual restrictions that one needs depend on the application domain. It is useful to define $I_S|x$ when $x \in G^m$ as the subset of I_S that contains all the restrictions that involve x (as described in §5.5). It is also convenient in what follows to define a two place predicate, 'restriction' as below (here G is as above).

- $M \models_{w,t} \text{restriction}(G, r)$ iff $r \in G.I_S \cup G.I_R$

5.5 Performing Strategies

Now we have all the formal machinery needed to get to the important definitions of this paper. I describe them informally and then state them formally. A group 'performs' a strategy along a particular scenario iff its members 'perform' their parts of the strategy along that scenario, while interacting in such a way as to 'comply-with' all the restrictions imposed by the structure of the group. An individual agent 'performs' a strategy along a particular scenario iff he can first 'achieve' the 'current' part of that strategy, and then 'perform' the 'rest' of his strategy, both along that scenario. Thus this definition first recurses on the structure of groups, and then on the definition of strategies (using 'current' and 'rest' as described in §5.3). The 'current' of a strategy must be of the form 'do(A)' or 'wait(A)' and is 'achieved' by doing an appropriate sequence of actions—this is where the reactive component of the architecture applies. A group's actions 'comply-with' all the group structure restrictions if all the actions called for by those restrictions are done; i.e., commands are obeyed, and temporal order constraints on the actions of different members are not violated.

With these definitions in hand, I turn to the conditions that the successful performing of a strategy by a group would bring about. These are captured by the predicate 'leads-to' and are the primary external criterion for distinguishing among the intentions of different groups. One important difference between the approach of this paper (as it is embodied in the definition of 'performs') and that of [Singh, 1991a] (as it is embodied in the definition of 'follows' given there) is that the former considers all sequences of actions that the group can do to succeed with its strategy, while the latter considers only those scenarios over which the group actually *forces* the success of its strategy. This provides the basis of the difference between intentions and ability—the key idea being that neither should entail the other.

Each scenario in the model determines a sequence of actions by a group (as described in §5.1), but a sequence may occur over any of several scenarios. In the following definitions, $G = \langle\langle x_1, \ldots, x_n\rangle, I_S, I_R\rangle$ as before; s, the sequence of actions for G, is the parallel composition of sequences of actions of G's members. G 'performs' a strategy, Y, over a scenario iff all its members 'perform' their parts of Y over that scenario while complying with all its restrictions. The members of G, i.e., the x_i, could themselves be groups. The predicate, 'complies-with,' is meant to accommodate the restrictions on the interactions among agents, and has to be defined for each kind of restriction in $I_S \cup I_R$. I come back to it later in this section.

- $M \models_{w,t}$ performs(G, s, t', Y) iff $(\exists S : S \in \mathbf{S}_{w,t} \wedge \langle S, t, t'\rangle \in [\![s]\!]) \wedge (\forall r : r \in (I_S \cup I_R)$ $\rightarrow$ complies-with$(r, t, t')) \wedge (\forall i : i \in [1, \ldots, n] \rightarrow M \models_{w,t}$ performs$(x_i, s'_i, t', Y_i))$

In the following definitions, x is a single agent; s a sequence of actions of x, s' a prefix of sequence s, and s'' the rest of it. Further, ϵ denotes the empty sequence, and t' denotes a time where s ends, if begun at t. As formally defined later in this section, x 'achieves' the 'current' part of his strategy as s' occurs, and the remaining part of it as s'' occurs. Since this definition may be invoked when x is a member of a group, x may take less time to perform Y than allowed, i.e., till t'', rather than till t'. The base case ('skip') is trivial.

- $M \models_{w,t}$ performs$(x, \epsilon, t,$ skip$)$
- $M \models_{w,t}$ performs(x, s, t', Y) iff $(\exists S, t'' : S \in \mathbf{S}_{w,t} \wedge t \leq t'' \leq t' \wedge \langle S, t, t''\rangle \in [\![s]\!] \wedge$ $(\exists s', s'', t''' : t \leq t''' \leq t'' \wedge s' \neq \epsilon \wedge s = s' \circ s'' \wedge M \models_{w,t}$ achieves$(x_i, s'_i, t''',$ current$(Y_i)) \wedge$ $M \models_{w,t'''}$ performs$(x, s'', t'',$ rest$(Y))))$

Now the predicate 'leads-to' may be defined as follows. For a group, a strategy leads to a condition iff that condition obtains over all scenarios over which that group can 'perform' that strategy. If the strategy cannot be 'performed,' then this definition applies vacuously. Here x may be a group or an individual agent.

- $M \models_{w,t}$ leads-to(x, Y, p) iff $(\forall S : (S \in \mathbf{S}_{w,t} \wedge (\exists t' : (\exists s : M \models_{w,t}$ performs$(x, s, t', Y))))$ $\rightarrow (\exists t'' : t'' \in S \wedge M \models_{w,t''} p))$

In the definition of ‘performs,’ since ‘skip’ is the empty strategy, ‘achieves’ is invoked only for cases (1) and (2), and is defined for them below. Here I assume that *Constr* is the conjunction of background constraints (e.g., to not run into a car, and to not die) that must never be violated. ‘Achieves’ captures the inherent reactivity of real-life agents—typically, a sequence of actions that just includes what a strategy prescribes would violate some constraint, or fail to achieve the relevant condition. ‘Seq’ achieves ‘do(A)’ along a scenario specified by t' iff the group achieves A in doing ‘seq’ over that scenario and the background constraints are preserved as ‘seq’ is done. This allows agents to over act, i.e., to do more than A requires (because of their habits, perhaps). ‘Seq’ achieves ‘wait(A)’ along a scenario specified by t' iff it terminates just when A occurs, and the background constraints are preserved as it is done.

- $M \models_{w,t}$ achieves(G, seq, t', do(A)) iff $(\forall S : S \in \mathbf{S}_{w,t} \wedge (\exists t'' : t'' \leq t' \wedge \langle S,t,t''\rangle \in [\![\text{seq}]\!]) \rightarrow (\exists t'' : \langle S,t,t''\rangle \in [\![\text{seq}]\!] \wedge (\forall t''' : t \leq t''' \leq t'' \rightarrow M \models_{w,t'''} \mathit{Constr}) \wedge (\exists t''' : t \leq t''' \leq t'' : M \models_{w,t'''} A)))$
- $M \models_{w,t}$ achieves(G, seq, t', wait(A)) iff $(\forall S : S \in \mathbf{S}_{w,t} \wedge (\exists t'' : t'' \leq t' \wedge \langle S,t,t''\rangle \in [\![\text{seq}]\!]) \rightarrow (\exists t'' : \langle S,t,t''\rangle \in [\![\text{seq}]\!] \wedge (M \models_{w,t''} A) \wedge (\forall t''' : t \leq t''' \leq t'' \rightarrow M \models_{w,t'''} \mathit{Constr})))$

The predicate, ‘complies-with,’ is meant to accommodate the restrictions on the interactions among agents, and has to be defined for each kind of restriction in $I_S \cup I_R$. The kinds of restrictions that would be of interest depend on the domain that is being modeled. I consider two important classes of restrictions that occur in many application domains. The first kind involve the temporal ordering of the substrategies of the members of G (and is then in I_S), or the ordering of their actions (and is then in I_R). Conditions that involve the relative ordering of actions of members can then be handled in the formalism in the same manner as the ordering of substrategies, though their implementational import is quite different. For I_S, I define “labels” as propositions that are true exactly when the substrategy they name has been successfully followed; for I_R, they would just denote the completion of some basic action or the achievement of some appropriate condition. This allows us to express temporal conditions such as ‘l_p before l_q,’ which states that l_p always occurs before l_q. Other conditions can be similarly handled. Note that if l_p or l_q names a substrategy of member x then ‘l_p before $l'_q \in I_S|x$.

- $M \models_{w,t}$ complies-with(‘l_p before l_q’, t, t') iff $(\forall S : S \in \mathbf{S}_{w,t} \wedge t, t' \in S \rightarrow (\forall t'' : t \leq t'' \leq t' \wedge M \models_{w,t''} l_q \rightarrow (\exists t''' : t \leq t''' < t'' \wedge M \models_{w,t'''} l_p)))$

Other kinds of restrictions in I_S correspond to the different illocutionary acts that members are able to perform. Different restrictions would state that a member’s commands to another would always be obeyed, or that a member’s request to another would be accepted if certain conditions obtain, and so on. I consider only commands here; other kinds of illocutionary acts may be included and given the exact semantics ones domain requires one gives them. Commands can be given only by agents to their subordinates: I extend I_S to include statements of hierarchy: ‘x_i commands x_j’ (in other words, x_i is the master and x_j the slave). Now let “command(agent, condition)” be an action that a member’s (in this case, the master's) strategy may require to be 'done.' A command restriction is 'complied-with' over a time interval, if all commands by superiors (as defined by that restriction) to their subordinates are obeyed within that interval. Then,

- $M \models_{w,t}$ complies-with('x_i commands x_j', t, t') iff $(\forall S : S \in \mathbf{S}_{w,t} \wedge t, t' \in S \rightarrow (\forall t_1, t_2 : t \leq t_1 \leq t_2 \leq t' \wedge \langle S, t_1, t_2 \rangle \in [\![\text{do}(\text{command}(x_j, A))]\!]^{x_i} \rightarrow \ (\exists t_3 : t_2 \leq t_3 \leq t' \wedge M \models_{w,t_3} A)))$

6 Group Intentions

As for any folk concept that we try to formalize, several different senses of group intention can be defined. Each of these senses corresponds to a slightly different concept with its own trade-offs and applicability. Using the formal model already defined, I now present a general framework in which several different senses of intentions may be formalized. However, since endless variations are possible in principle, I begin with a simple definition, explain its ramifications, and apply it to a detailed example. I then present some of those definitions that are the most reasonable from the point of view of AI.

Given the preceding analysis of strategies that groups may have and the ways in which those strategies may be performed, one can come up with a fairly simple and general definition of intentions. This definition simply states that *a group intends all the necessary consequences of its performing its strategy.* Note that only the consequences of the successful performing of the strategy are included. There is no guarantee that a given strategy will in fact be successfully performed. Despite its simplicity, this definition incorporates three subtle features that make it quite powerful:

- This definition applies uniformly to single agents and complex groups. Thus intentions of agents and intentions of groups are the same kind of entity. This is an improvement over the traditional theories' definitions described in §3, which define the intentions of individuals and groups to be concepts of different categories.

- This definition considers as intentions only the *necessary* consequences of the performing of the group's strategy. This is important because we clearly do not want to claim that a group intends even the merely contingent consequences of its performing its strategy. E.g., an agent with a strategy for loading paper in a printer will have to pick some ream of paper or the other, but cannot be said to have had the intention of picking the one it finally picked. This is because it could just as well have picked the other one. On the other hand, if it was forced to pull out the paper-feeder tray, we can say from the external perspective that it must have intended to pull it out. Note that the agent might have done that action *intentionally*, but need not have had the prior intention to pick up that very ream—this distinction is described by Bratman [1987, p. 119].

- This definition considers the *performing* of a strategy by a group. Thus it accommodates both (a) the inherent reactivity that agents groups must exhibit in performing their strategies and (b) the impact that the social structure of a group has on its performing different strategies.

> Note also that this definition uses the necessary consequences of performing a strategy. Thus a group's limitations of ability to do certain actions, as well as its habits are taken into account. For example, consider a group of two agents engaged in adding paper to a printer. Let it be a habit of this group that one of the agents picks up a fresh ream of paper after the other has pulled out the tray from the printer. It would be all right externally to attribute to this group not only the intention of adding paper, but also the intention of picking up a new ream after the tray is pulled out. For a group that did not have the abovementioned habit, the same strategy would not yield the latter intention.

In more detail, the above definition, which will also serve as the "core" definition in this paper is simply that a group G intends^e p iff G 'has' a strategy that 'leads-to' p (here the superscript ^{e} is meant to remind us that this is an "external" definition). And formally, we have

- $M \models_{w,t} \text{intends}^e(G, p)$ iff $(\exists Y : M \models_{w,t} \text{has}(G, Y) \wedge M \models_{w,t} \text{leads-to}(G, Y, p))$

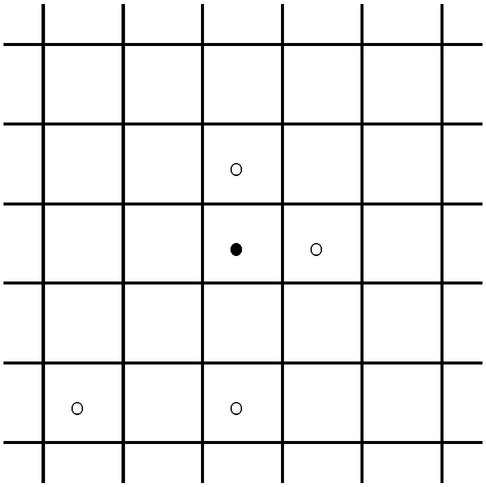
The predicate 'leads-to' was formally defined in §5.5; informally, it means that condition p is made true at some point on all scenarios where group G performs strategy Y; i.e., if G correctly performs Y, p would necessarily occur. The predicate 'has' captures the requirement that the strategy attributed to G be, in fact, a strategy that G has. In this way, this definition gives equal importance to reactive actions (through 'leads-to') and to internal states (through 'has').

Since groups may be nested, the definition of 'has' must be recursive in the structure of groups. For a group to 'have' a strategy its members must 'have' their parts of it. The base case (for an individual, x) comes directly from the model; $\mathbf{C^Y}$ is a part of the interpretation (as defined in §5.1):

- $M \models_{w,t} \text{has}(x, Y)$ iff $Y \in \mathbf{C^Y}_{w,t}(x)$
- $M \models_{w,t} \text{has}(G, Y)$ iff $(\forall i : x_i \in G^m \rightarrow \ \text{has}(x_i, Y_i))$

6.1 Example: The Pursuit Problem

The theory as developed so far is simple, but is nevertheless quite useful in modeling multiagent systems. I now apply it to the analysis of a well-known problem in multiagent systems in AI: the pursuit problem. This problem was introduced by Benda *et al.* [1986] and has been extensively studied by others [Durfee & Montgomery, 1989; Stephens & Merx, 1989]. This problem has been analyzed from several different perspectives. Here my aim is simply to analyze the intentions of the team of Blue agents in terms of the intentions of the individual Blue agents. We are given a finite two-dimensional grid of points (see Figure 2). Each point may be occupied by either an agent called 'Red' (the "adversary") *or* up to four 'Blue' agents. At each cycle, each agent can stay in its location or move one square up, down, left, or right. The pursuit starts in some arbitrary configuration and ends in either the Blue agents winning (when they occupy the four locations surrounding Red) or losing



(• = Red, ∘ = Blue)

Figure 2: The Pursuit Problem: an example configuration

(if Red gets to the edge of the grid). For simplicity, also assume that Red always moves to avoid abutting a Blue agent, if he has such a move available.

Let the Blue agents be notated as B^1 through B^4, Red as R, and any of the five as A. Let A_x denote the x coordinate of A's location, and A_y the y coordinate. Initially, let the agents be specified to get on different sides of R, e.g., B^1 above it, B^2 to its right, B^3 below it, and B^4 to its left.

At the most abstract level of description, each of the Blue agents could be assigned a strategy (i.e., be said to 'have' a strategy) that just requires it to go its proper slot; e.g., Y_1 could just be 'do(get-above-R)'; the other strategies would be analogous. This specification assumes that lower layers of the design are available so that the required actions may be done. These layers would also ensure that no constraint is violated (e.g., collision with Red is avoided). More importantly, each of the Blue agents could be said to have some intentions on the basis of their strategies. E.g.,

- $M \models_{w,t} \text{intends}^e(B^1, ((B^1_x, B^1_y) = (R_x, R_y + 1)))$

That is, B^1 has the intention to occupy the location just above R; the other Blue agents' intentions are analogous. Now the group as a whole has a strategy whose substrategies are the agents' strategies. No restrictions have been imposed on this group. Therefore, for the group of Blue agents, G, we have

- $M \models_{w,t} \text{intends}^e(G, \mathsf{P}((B^1_x, B^1_y) = (R_x, R_y{+}1)) \wedge \mathsf{P}((B^2_x, B^2_y) = (R_x{+}1, R_y)) \wedge \mathsf{P}((B^3_x, B^3_y) = (R_x, R_y - 1)) \wedge \mathsf{P}((B^4_x, B^4_y) = (R_x - 1, R_y)))$

That is, the group as a whole intends that B^1 occupy the location above Red, B^2 the one to its right, and so on, in some arbitrary temporal order. In other words, the condition intended by the group as a whole is not that it *win* (which requires that the Blue agents be in the appropriate positions at the same time). This provides a simple example of how a formal theory of intentions may be employed, namely, in the validation of designs. If what the designer really wants is that the group intend to win, he must arrange for the member agents to have different intentions than in the previous case or structure the group (i.e., constrain the members' interactions) appropriately. One possible solution for the designer is to have the agents each intend a more complex condition, e.g., to achieve the condition of being always in the appropriate position relative to Red; e.g.,

- $M \models_{w,t} \text{intends}^e(B^1, \mathsf{G}((B^1_x, B^1_y) = (R_x, R_y + 1)))$

Now it follows that for the group of Blue agents, G, we have

- $M \models_{w,t} \text{intends}^e(G, ((B^1_x, B^1_y) = (R_x, R_y + 1)) \wedge ((B^2_x, B^2_y) = (R_x + 1, R_y)) \wedge ((B^3_x, B^3_y) = (R_x, R_y - 1)) \wedge ((B^4_x, B^4_y) = (R_x - 1, R_y)))$

In other words, the group intends to *win*. It is obvious from the construction that none of the Blue agents intends anything beyond occupying a certain position relative to Red. The intention of the group derives simply from the combination of the members' intentions.

The strategies given above for the Blue agents are so abstract that they provide no clue as to how the agents must act. At the next level of detail, we might assign strategies that are more detailed descriptions of the agents' actions. Let Y_1 be

```
while $\neg \text{over}_1$ do
    if $B^1_y > R_y + 1$ then do(move-down)
    else if $B^1_x < R_x$ then do(move-right)
        else if $B^1_x > R_x$ then do(move-left)
            else if $B^1_y < R_y$ then do(move-up) else skip
```

The other Blue agents' strategies are analogous. Here I assume that 'over_i' is true when B^i abuts Red on the appropriate side for sufficiently long. As before, I also assume that the lower layers of the architecture ensure that no constraint is violated. The above strategies provide a more detailed specification of what a Blue agent would do, and thus of its internal architecture. They not only validate the claim that each agent has an intention to occupy a certain position relative to Red but also validate certain others; e.g., if B^1 is far above Red and Red does not move up then B^1 intends to move downwards before moving laterally. This could not have been concluded in the previous case, where the strategies were specified in somewhat less detail. However, again the group as a whole has the intention to achieve a *win*.

In both these cases intentions were attributed to the group of Blue agents quite independently of whether the individual agents knew of them or planned to achieve them. Indeed, the group itself was described independently of whether its members knew they participated in it. A more elaborate system in which the agents took on goals depending on their location could also be described in this framework: their strategies would have to be more complex; e.g., they could each decide which location to go to depending on some globally optimal assignment. At greater levels of detail, the strategies assigned to the agents could include the aspects of interaction required for them to adopt non-conflicting goals. For concreteness, consider a system where B^1 becomes the controller, and the other Blue agents its slaves. Now Y_1 could begin by first making an assignment of the locations surrounding Red to different Blue agents and then *command* the other Blue agents to go there. Y_2, Y_3 and Y_4 could then simply be strategies to go to any specified location about Red. The group would impose the strategic restriction that all commands from B^1 are obeyed (as described in §5.5). Again, the intentions ascribed previously could be ascribed, along with some more complex ones, e.g., about the specific order in which actions are done by different agents.

6.2 Further Variations

The sense in which a group may be said to 'have' a strategy determines the sense of group intention being defined. In the most basic sense of 'has,' which was considered above, the members need not even be aware that they are a part of a group. In other words, this is the most external sense of group intentions in the present framework—the group is defined entirely from without by an objective observer, as it were. This ascription is exactly the one we want when speaking of certain kinds of non-introspective agents, e.g., ant colonies.

However, it is sometimes desirable to ascribe an intention to a group only when its members participate more actively in it, or are aware to a larger degree of their role in the group and of the goal it is trying to achieve. This may be required not just to give a truer account of several multiagent systems, but also for certain practical reasons. A (cooperative) group whose members were aware of it would, in general, be more robust; i.e., it could achieve its goals under a wider variety of conditions. E.g., a Blue agent may defer achieving its own goal if it can instead prevent Red from escaping (and thus from defeating the entire group). Secondly, architectural considerations on the following of strategies (e.g., the perceptual hardware required) can be more easily described when the members are aware of their group. E.g., a Blue agent who expected some information over a channel may keep it free; one who expected a neighboring agent to receive some important information from someone else could put off communicating with it. We can also limit the ascription of intentions (and beliefs) to those conditions that we have architectural reasons to believe are those that the agent could directly perceive or influence; i.e., while an agent would not be able to distinguish between logically equivalent conditions, he might be more directly causally connected to one of them than the other; e.g., a clam might be directly connected to the tide being high rather than the moon being high in the sky (assuming they are logically equivalent conditions).

In defining these stronger senses of group intentions, we can require that each member of a group be more introspective about it. Thus each member can be required to be aware of its strategic interactions with other members, or be aware of the strategic interactions of all members, or be aware of the strategies of all members, and so on. These successive requirements make the members know more and more about their group. I write has_k when condition k below is assumed (MB means "mutually believe"):

- $M \models_{w,t} \text{has}(G, Y)$ iff $(\forall i : x_i \in G^m \rightarrow\ \text{has}(x_i, Y_i)) \wedge$

1. $(\forall i : x_i \in G^m \rightarrow (\forall r : r \in (I_S|x_i) \rightarrow\ \text{believes}(x_i,\ \text{restriction}(G, r))))$

2. $(\forall i : x_i \in G^m \rightarrow (\forall r : r \in I_S \rightarrow \text{ believes}(x_i, \text{ restriction}(G, r))))$
3. $(\forall i, j : x_i, x_j \in G^m \rightarrow \text{ believes}(x_i, \text{ has}(x_j, Y_j)))$
4. $(\forall i : x_i \in G^m \rightarrow \text{ MB}(x_1, \ldots, x_n; \text{ has}(x_i, Y_i)))$

The predicate 'believes' needs to be defined, at least for the predicates over which it is used above. Here too, since groups may be nested, the definition must be recursive in the structure of groups. For a group, G, to 'believe' that r is an interaction (e.g., one in which G as a whole participates as a member (in the group that contains it)), at least some appropriate components of G that would take care of r would need to 'believe' so. However, what these components are is not obvious from the strategies and structure of G. Since I am focusing on intentions in this paper, I simply define 'believes' to hold for all members of G.

- $M \models_{w,t}$ believes(x, p) iff $p \in \mathbf{C^B}_{w,t}(x)$
- $M \models_{w,t}$ believes(G, p) iff $(\forall i : x_i \in G^m \rightarrow \text{ believes}(x_i, p))$

The use of the predicates introduced above, MB and 'believes,' makes this theory applicable to robust groups (of more than minimally intelligent agents). These predicates help us assign strategies that are more abstract and schematic. While for sufficiently detailed strategies these predicates would not be needed at all, they are quite often useful in practice. They must both be given definitions in terms of the groups' actions and internal structure (at least for the instances over which they are used) for the entire theory to continue to be objective. This task is not attempted in this paper. For the single agent case, and in the absence of explicit consideration of intentions, this problem has been addressed by others [Stalnaker, 1984]; further work is needed to make the connection with internal structure explicit and to explore the relativity of beliefs to intention ascriptions. The addition of new constraints yields variations of this theory that appear close to the traditional theories in terms of the logical axioms validated. However, even then this theory, because of its consideration of the internal structure of groups, provides a better motivation for those axioms than the traditional theories.

A property of the above definition of intends^e is that for all the different senses of 'has,' it validates consequential closure; i.e., if G intends^e p and p entails q, then G also intends^e q.[1] Since a strategy that 'leads-to' p automatically 'leads-to' q, this inference is acceptable in an external analysis (when we ignore the mental state of the agents except to the extent that this is reflected in their behavior). However, it is not always appropriate, since we often want to talk of intentions in a sense that gives importance to a psychological interpretation. In such a sense, an intention of an agent may have consequences that are not intended [Bratman, 1987, p. 140].

This inference can be prevented by including a direct notion of what agents (and groups) intend. The new predicate, 'intends-for,' tells us which strategy an agent has and what condition that strategy is meant to achieve. The second part of the meaning of 'intends-for' acts like a syntactic filter here and captures the architectural limitations of agents, as described earlier in this section. Fagin and Halpern have used this idea in the context of

[1] Note that if p only contingently implies q (i.e., if p does not entail q) then the resulting inference would be invalidated by the current definition (as it should be).

beliefs [1988]. The ideas of this paper can also be combined with the approach of [Singh & Asher, 1992], but since that approach is quite complex and largely orthogonal to the issues addressed in this paper, that is not done here. For simplicity, I assume that in all contexts, e.g., in the context of 'believes,' 'intends-for(G, Y, p)' entails 'has(G, Y).' I also assume that in all contexts 'intends-for(G, Y, p)' entails 'believes(x, leads-to(Y, p)).' This helps simplify the relationship with Grosz and Sidner's theory.

Now we can define intends$^i(G, p)$ as follows (here the superscript ^{i} is for "internal"):

- $M \models_{w,t}$ intends$^i(G, p)$ iff $(\exists Y : M \models_{w,t}$ has$(G, Y) \wedge M \models_{w,t}$ leads-to$(G, Y, p) \wedge M \models_{w,t}$ intends-for$(G, Y, p))$

The conjunct involving 'has' is redundant but is included to use the preceding classification of has$_k$'s. The predicate 'intends-for' too has to be given a recursive definition in the structure of groups. The base case, when x is an individual, is ($\mathbf{C^I}$ is a part of the interpretation)

- $M \models_{w,t}$ intends-for(x, Y, p) iff $\langle Y, p\rangle \in \mathbf{C^I}_{w,t}(x)$

For groups, 'intends-for' may be defined in several ways, each sufficient for blocking consequential closure. Let group $G = \langle\langle x_1, \ldots, x_n\rangle, I_S, I_R\rangle$ as always.

1. $M \models_{w,t}$ intends-for$_1(G, Y, p)$ iff $(\exists i : x_i \in G^m \wedge$ intends-for$_1(x_i, Y_i, p))$
2. $M \models_{w,t}$ intends-for$_2(G, Y, p)$ iff $(\forall i : x_i \in G^m \rightarrow$ intends-for$_2(x_i, Y_i, p))$
3. $M \models_{w,t}$ intends-for$_3(G, Y, p)$ iff $(\forall i : x_i \in G^m \rightarrow$ MB$(x_1, \ldots, x_n;$ (intends-for$_3(x_i, Y_i, p) \wedge G = \langle\langle x_1, \ldots, x_n\rangle, I_S, I_R\rangle)))$

By the above assumption, intends-for$_3$ entails has$_3$ and has$_4$. The definition of 'intends' resulting from intends-for$_3$ corresponds closely to that of Grosz and Sidner.

7 Some Formal Consequences

I now enumerate some formal consequences of the above definitions. In theoretical work such as the one described here, the negative consequences (namely, the implications that do not hold) are as important as the positive ones. Accordingly, they are included below. These consequences vary for the different sense of intentions and serve to distinguish them from each other in a principled manner. For brevity, 'intends' is used for both intends^e and intends^i. Validity is indicated by '$\models$,' and invalidity by '$\not\models$.'

1. **Singleton Group Formation (SGF):**

   $\models$ intends$(x, p) \leftrightarrow$ intends$(\langle\langle x\rangle, \emptyset, \emptyset\rangle, p)$

   This inference is valid when has$_1$ and has$_2$ are used in the definition of 'intends.' Unlike in the theories of Grosz and Sidner [1990], and Cohen and Levesque [1988], here the group containing just x (and therefore having no restrictions), namely $\langle\langle x\rangle, \emptyset, \emptyset\rangle$, has the same intentions as x. This makes sense because simply calling an agent a group should not change anything.

For example, let Bill (as an agent) intend to go to Italy. Then the group formed out of Bill alone and with no restrictions 'has' (i.e., has_1 or has_2) a group strategy whose only component strategy is Bill's strategy (strategies are assigned to agents by the interpretation, as defined in §5.1). This strategy is performed by all sequences by which Bill's strategy is performed. So the group has the same intentions as Bill. When the other senses of 'has' are used, then the group need not 'have' the strategy whose only component is Bill's strategy since the unique member of the group (i.e., Bill himself) might not believe that he 'has' the strategy that he in fact 'has.'

2. **Temporal Coherence (TC):**

   $\models [\text{intends}^e(x,p) \wedge \text{intends}^e(y,q) \wedge (x,y \in G^m)] \rightarrow \text{intends}^e(G,[\mathsf{P}p \wedge \mathsf{P}q])$

   P stands for "sometimes in the past." This inference is valid for has_1 and has_2. If two members of a group have strategies that lead-to p and q, respectively, then there is a group strategy that includes those strategies, and leads-to p and q in some arbitrary order. This is because on all the scenarios where p occurs and q occurs, they must occur in some temporal order (or simultaneously). Note that this may hold vacuously because the appropriate group strategy may be impossible to perform successfully (perhaps due to the interactions required). Therefore, if we externally ascribe intentions for two conditions to a group, we must also ascribe to it the intention to achieve those two conditions in some order. E.g., if an ant colony intends to make one new cavern, and to make another new cavern, it intends to make them in some temporal order.

   This inference is invalid for intends^i since it is possible that no member 'intends-for' the complex condition on the right hand side. This corresponds to the intuition that the group might not internally intend a consequence of its intentions, because it does not internally know of it.

3. **Consequential Closure (CC):**

   $\models \text{intends}^e(G,p) \wedge \Box(p \rightarrow q) \rightarrow \text{intends}^e(G,q)$

   $\Box(p \rightarrow q)$ means that p entails q; i.e., at every world and time, p implies q. The given inference is valid for all the has_k defined in the paper because of the definition of 'leads-to'—G's strategy 'leads-to' p if and only if on all scenarios where that strategy is performed, p occurs. But since p entails q, wherever p occurs, q must occur also. Thus G's strategy 'leads-to' q as well. Hence, irrespective of the sense has_k in which G 'has' that strategy, it intends q. E.g., if an ant colony intends to make a new cavern and making a new cavern entails making a new arch, then it intends to make a new arch as well.

4. **Strong Upward Closure Through Group Formation (SUC):**

   $\models (\exists x : x \in G^m \wedge \text{intends}^e(x,p)) \rightarrow \text{intends}^e(G,p)$

   This inference is valid for has_1 and has_2. It is valid since x's strategy is a component of the group's strategy in the sense of has_1 and has_2. Any scenario over which the group strategy is performed must be a scenario over which x's strategy is also performed. But, by the definition of intends^e, p occurs on all the scenarios where x's strategy is performed. Therefore, p occurs on all scenarios where the group's strategy is performed. Hence, again using the definition of intends^e, the group intends q. By a similar argument, this inference is also valid for intends^i with intends-for$_1$. It is not valid for intends^i with intends-for$_2$ or intends-for$_3$ because in those cases (as for intends^e with has$_3$ or has$_4$), the group (due to the belief requirements) might not have a strategy whose component is the strategy of member x. An example of this inference is the following. If a platoon of an army brigade intends to control a major bridge, then the army brigade intends to control that bridge.

5. **Weak Upward Closure Through Group Formation (WUC):**

   $\models (\forall x : x \in G^m \rightarrow \text{intends}^e(x,p)) \rightarrow \text{intends}^e(G,p)$

   This inference is valid for intends^e with has$_1$ and has$_2$ for quite the same reasons as given above. It is also valid for intends^i with intends-for$_1$ or intends-for$_2$ because the antecedent condition requires that all the members of the group have strategies that 'lead-to' p. Thus the group has a strategy whose components are the strategies of the members. Since each of those strategies 'lead-to' p, the strategy of the group also 'leads-to' p. This inference is not valid for intends^i with intends-for$_3$, because the members do not necessarily have a mutual belief that they have the strategies that they in fact have. So in this sense, the group might not have the appropriate strategy needed to validate the inference. An example of this inference is the following. If several ants in a colony intend to repel an invader, then they as a group intend to repel the invader. Also, if all the prisoners in a jail intend (independently) to poison the guards, then they as a group intend to do so too.

6. **Weak Downward Closure Through Group Formation (WDC):**

   $\not\models \text{intends}(G,p) \rightarrow (\exists x : x \in G^m \wedge \text{intends}(x,p))$

   This inference says that the intentions of a group may emerge from the intentions of its members; i.e., it is possible that no member of a group has the same intention as the group itself. It is not valid formally since the group's intention may be due to a combination of its members' intentions; e.g., in item 2 above, the intention of the group is a temporally complex condition derived from the intentions of its members. Another example is of this would be a set of people each intending to move himself into a train at a subway station. The intention of the set as a group is to move the *entire* set of people into the train, which is something that none might have intended (and a few, who prefer a lot of breathing space, might have hoped would not happen).

7. **Strong Downward Closure Through Group Formation (SDC):**

   $\not\models \text{intends}(G,p) \rightarrow (\forall i : x_i \in G^m \rightarrow \text{MB}(x_1,\ldots,x_n; \text{intends}(G,p)))$

   This inference is invalid for the same reasons as inference 6—the group's intention may be a combination of its member's strategies as in, e.g., item 2. However, it is valid for intends^i with intends-for$_3$ (as in Grosz and Sidner's theory). This is because intends-for$_3$ explicitly requires a mutual belief among the members of the group about the identity of the group and about the strategies of its members. Jointly, these yield a mutual belief about the strategy of the group, which combines with the mutual belief involving 'leads-to' to a mutual belief in the intention of the group (assuming the agents are smart enough to do the necessary computations). An example of this inference is when a few prison inmates, who trust each other perfectly and mutually know that they do, are considered as a group. Therefore, if the group intends to poison the guards and let every member escape, then the members mutually believe that the group intends so.

The consequences enumerated above involve several important theorems and non-theorems of the theory of this paper. They help distinguish between the different senses of intentions on the major dimensions and help relate our model-theoretic intuitions (i.e., the ones about the structure of groups and their strategies) to our proof-theoretic intuitions (i.e., the ones about which inferences are valid in which kind of situation).

8 Conclusions

The theory presented in this paper refines and formalizes some intuitions about group intention, especially as that concept may be used in the modeling the behavior of complex multiagent systems. It attempts to ground this concept in terms of (1) the actions done by the members of a group of agents, and (2) their social structure, as it emerges from their interactions. It allows nested groups, and provides a framework in which several useful senses of group intention can be formalized; in the general case, this theory imposes far weaker knowledge requirements on agents than the traditional theories of intentions in multiagent systems. The notion of strategies, and of their being performed reactively by individual agents and groups, provides a useful and perspicuous link to the theory of group ability. The theory as developed so far captures many interesting senses of intentions. Further work, however, is needed to extend this framework to other kinds of groups, and to develop an account of deliberation by a group about the intentions it has or might adopt.

References

[Agre & Chapman, 1987] Agre, Philip and Chapman, David; 1987. Pengi: An implementation of a theory of activity. In *AAAI*. 268–272.

[Austin, 1962] Austin, John L.; 1962. *How to do Things with Words*. Clarendon, Oxford, UK.

[Barwise & Perry, 1983] Barwise, Jon and Perry, John; 1983. *Situations and Attitudes*. MIT Press, Cambridge, MA.

[Benda *et al.*, 1986] Benda, Miroslav; Jaganathan, V.; and Dodhiawala, Rajendra; 1986. On optimal cooperation of knowledge sources. Technical report, Boeing Advanced Technology Center, Boeing Computer Services, Seattle, WA.

[Bratman, 1987] Bratman, Michael E.; 1987. *Intention, Plans, and Practical Reason.* Harvard University Press, Cambridge, MA.

[Chellas, 1980] Chellas, Brian F.; 1980. *Modal Logic.* Cambridge University Press, New York, NY.

[Cohen & Levesque, 1988] Cohen, Philip R. and Levesque, Hector J.; 1988. On acting together: Joint intentions for intelligent agents. In *Workshop on Distributed Artificial Intelligence.*

[Davis & Smith, 1983] Davis, Randall and Smith, Reid G.; 1983. Negotiation as a metaphor for distributed problem solving. *Artificial Intelligence* 20:63–109. Reprinted in *Readings in Distributed Artificial Intelligence*, A. H. Bond and L. Gasser, eds., Morgan Kaufmann, 1988.

[Demazeau & Müller, 1990] Demazeau, Y. and Müller, J-P., editors. *Decentralized Artificial Intelligence*, Amsterdam, Holland. Elsevier Science Publishers B.V. / North-Holland.

[Dennett, 1987] Dennett, Daniel C.; 1987. *The Intentional Stance.* MIT Press, Cambridge, MA.

[Durfee & Montgomery, 1989] Durfee, Edmund H. and Montgomery, Thomas A.; 1989. MICE: a flexible testbed for intelligent coordination experiments. In *Proc. 9th Workshop on Distributed Artificial Intelligence.* 25–40.

[Emerson, 1990] Emerson, E. A.; 1990. Temporal and modal logic. In Leeuwen, J.van, editor, *Handbook of Theoretical Computer Science*, volume B. North-Holland Publishing Company, Amsterdam, The Netherlands.

[Fagin & Halpern, 1988] Fagin, Ronald and Halpern, Joseph Y.; 1988. Belief, awareness, and limited reasoning. *Artificial Intelligence* 34:39–76.

[Fischer & Immerman, 1986] Fischer, Michael J. and Immerman, Neil; 1986. Foundations of knowledge for distributed systems. In Halpern, Joseph Y., editor, *Theoretical Aspects of Reasoning About Knowledge.* 171–185.

[Fischer & Ladner, 1979] Fischer, Michael J. and Ladner, Richard E.; 1979. The propositional dynamic logic of regular programs. *Journal of Computer and System Sciences* 18(2):194–211.

[Georgeff, 1987] Georgeff, Michael P.; 1987. Planning. In Traub, J. F., editor, *Annual Review of Computer Science, Vol 2.* Annual Reviews Inc., Palo Alto, CA.

[Grosz & Sidner, 1990] Grosz, Barbara and Sidner, Candace; 1990. Plans for discourse. In Cohen, P.; Morgan, J.; and Pollack, M., editors, *SDF Benchmark Series: Intentions in Communication.* MIT Press, Cambridge, MA.

[Halpern & Moses, 1987] Halpern, Joseph Y. and Moses, Yoram O.; 1987. Knowledge and common knowledge in a distributed environment (revised version). Technical Report RJ 4421, IBM.

[Hamblin, 1987] Hamblin, C. L.; 1987. *Imperatives.* Basil Blackwell Ltd., Oxford, UK.

[Hewitt, 1988] Hewitt, Carl; 1988. Organizational knowledge processing. In *Workshop on Distributed Artificial Intelligence.*

[Konolige, 1982] Konolige, Kurt G.; 1982. A first-order formalism of knowledge and action for multi-agent planning. In Hayes, J. E.; Mitchie, D.; and Pao, Y., editors, *Machine Intelligence 10.* Ellis Horwood Ltd., Chichester, UK. 41–73.

[Kozen & Tiurzyn, 1990] Kozen, Dexter and Tiurzyn, Jerzy; 1990. Logics of program. In Leeuwen, J.van, editor, *Handbook of Theoretical Computer Science.* North-Holland Publishing Company, Amsterdam, The Netherlands.

[McArthur, 1988] McArthur, Gregory L.; 1988. Reasoning about knowledge and belief: A survey. *Computational Intelligence* 4:223–243.

[McCarthy & Hayes, 1969] McCarthy, J. and Hayes, P. J.; 1969. Some philosophical problems from the standpoint of artificial intelligence. In *Machine Intelligence 4.* American Elsevier.

[McCarthy, 1979] McCarthy, John; 1979. Ascribing mental qualities to machines. In Ringle, Martin, editor, *Philosophical Perspectives in Artificial Intelligence.* Harvester Press. Page nos. from a revised version, issued as a report in 1987.

[Rosenschein, 1985] Rosenschein, Stanley J.; 1985. Formal theories of knowledge in AI and robotics. *New Generation Computing* 3(4).

[Singh & Asher, 1992] Singh, Munindar P. and Asher, Nicholas M.; 1992. A logic of intentions and beliefs. *Journal of Philosophical Logic.* In press.

[Singh, 1990a] Singh, Munindar P.; 1990a. Group intentions. In *10th Workshop on Distributed Artificial Intelligence.*

[Singh, 1990b] Singh, Munindar P.; 1990b. Towards a theory of situated know-how. In *9th European Conference on Artificial Intelligence.*

[Singh, 1991a] Singh, Munindar P.; 1991a. Group ability and structure. In Demazeau, Y. and Müller, J.-P., editors, *Decentralized Artificial Intelligence, Volume 2.* Elsevier Science Publishers B.V. / North-Holland, Amsterdam, Holland. 127–145.

[Singh, 1991b] Singh, Munindar P.; 1991b. A logic of situated know-how. In *National Conference on Artificial Intelligence (AAAI).*

[Singh, 1991c] Singh, Munindar P.; 1991c. Towards a formal theory of communication for multiagent systems. In *International Joint Conference on Artificial Intelligence (IJCAI).*

[Stalnaker, 1984] Stalnaker, Robert C.; 1984. *Inquiry.* MIT Press, Cambridge, MA.

[Stephens & Merx, 1989] Stephens, Larry M. and Merx, Matthais; 1989. Agent organization as an effector of DAI system performance. In *9th Workshop on Distributed Artificial Intelligence.*