In *10th Workshop on Distributed Artificial Intelligence*, October 1990

Group Intentions

Munindar P. Singh*

Dept of Computer Sciences and Artificial Intelligence Lab
University of Texas MCC
Austin, TX 78712-1188 Austin, TX 78759
USA USA

msingh@cs.utexas.edu
msingh@mcc.com

Abstract

Groups of intelligent agents are studied in several subareas of 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 group of agents that analyzes the intentions of a group in terms of its internal organization, and the intentions of its members. 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.

*I am indebted to Michael Huhns and to three anonymous referees for comments.

1 Introduction

Several subareas of Distributed Artificial Intelligence (DAI) are concerned with *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 [6, 7, 13, 16, 18]. In recent work, I have developed a formal theory of the ability of a group (as ascribed objectively) that accounts for its internal structure [24] and, with Nicholas Asher, a theory of intentions [26]. Here I plan to relate these interests, and motivate and present a formal theory of the intentions of a group of agents.

This theory is meant to apply to groups of agents of 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. It 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 [13] and Cohen and Levesque [6], 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 in this paper is in order here. Folk psychological notions such as belief and intention provide powerful abstractions with which to specify and model the behavior of complex systems [8, 20]. However, for these concepts to be effectively used in DAI (both science and engineering), they must be given an objective grounding in terms of the architectures that different kinds of systems have and the actions they do. My aim will be 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 concept of intention. 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. This strategy is itself seen as a set of strategies of its members. Following [24], the structure of a group is captured in terms of the "strategic" and "reactive" interactions of its members as they follow their respective strategies. As a consequence, this theory, unlike [6, 13], does not require mutual beliefs among the members of a group. Mutual beliefs are impossible to achieve in most realistic scenarios; e.g., where communication delay is not bounded or channels not reliable [12, 14].

I ought to clarify at the outset the meanings of two terms I use often in this paper: "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. The approach I have adopted in this paper is external in spirit, though it is not behavioristic—it is sometimes useful to take the agents' viewpoint seriously to better understand the constraints on their architectures. In its basic form, the theory developed here is meant to be used *about* agents rather than *by* them.

In §2, I present some intuitions and observations about groups, 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 of §2. In §4, I review some relevant parts of [24] extending it as needed for the purposes of this paper. In §5, 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 are needed for different DAI applications. In §6, I describe some of the logical consequences of the theory presented here.

2 Groups

Traditional theories of multiagent intentions and action tend to ignore the *structure* of the group of agents being considered [6, 10, 13, 22], and 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). Traditional theories also 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 mutually belief that they share that goal. However, 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 the AI Lab has created two pieces of software, RAD and XNETS, 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 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 (e.g., a bureaucracy, an ant colony and 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 (although its chances of success may be greatly reduced). In other situations, 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.

Two observations, made in [24], are also in order here.

1. A group (e.g., a sports team) may be considered as a single unstruc-

tured monolithic agent from without; i.e., groups are "Hobbesian corporate persons," in Hamblin's term [15, pp. 60, 240]. The intention of a group of agents is the same kind 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.

2. Even though groups (e.g., teams or corporations) seem monolithic from without, they are not homogeneous: typically, the members of a group 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 Security Council with five distinguished members. Groups may have further structure due to nesting; e.g., large corporations and armies are groups whose immediate constituents are also groups.

Other 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 [1, 25]. A good theory would accommodate the idea of situated action, and would consider the interactions among a group's members as they emerge from collective action. This idea is motivated and used in several recent papers; e.g., [11, 17]. 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 wish 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 also gives an account of the psychological phenomena that connects them directly to behavior and architecture, rather than requiring them to be independently attributed.

3 The Traditional Theories

The most well-known AI theories of group intentions are the ones of Grosz and Sidner [13], and Cohen and Levesque [6]. 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.

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 very complex and is only partially described here. They define a joint persistent goal as a goal that a group of agents mutually believe is a goal of each of them and 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, one that is for the proposition that the agents achieve the given intention and until the intention is achieved mutually believe that they are about to achieve it.

I lack the space to discuss these definitions in greater detail. 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 the general case, mutual beliefs are impossible to establish using asynchronous communications [12, 14]. 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. Agents 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 *"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: by requiring mutual beliefs they, in effect, require that groups be homogeneous—this is because a mutual belief is symmetric with respect to all the agents who have it. In fact, these 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 [21, 23, 25]. 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 applies at all, applies only to groups in 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 discourses 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 most.

In brief, the objections to traditional 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 The Framework

I now turn to a description of the formal model, and 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.

4.1 The Formal Model

The formal model here is quite close to the ones in [25, 24]; 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.

The formal model is based on possible worlds. 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. The actions an agent can do can differ at each time—this allows for learning and forgetting, and the acquisition and loss of physical capabilities, and for changes in the environment

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 $N = \langle [\![]\!], \mathbf{B}, \mathbf{C^Y}, \mathbf{C^B}, \mathbf{C^I} \rangle$ an interpretation. Here $\mathbf{W}$ is a set of possible worlds; $\mathbf{T}$ is a set of possible times; $<$ is partial order on $\mathbf{T}$; $\mathbf{I}$ is the class of domains of the various possible worlds; $\mathbf{A}$, the class of agents in different possible worlds, is a subclass of $\mathbf{I}$; $\mathbf{U}$ is the class of basic actions; as described below, $[\![\,]\!]$ assigns intensions to predicates and actions. $\mathbf{B}$ is the class of functions assigning basic actions 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 §5). Requirements, such as the persistence of agents with their strategies may be added if necessary. $\mathbf{C^B}$ is the class of functions assigning conditions of formulae to the agents at different worlds and times. These conditions are believed by the agents. $\mathbf{C^I}$ is the class of functions assigning pairs of strategies and conditions to the agents at different worlds and times. Intuitively, each 'intends-for' this strategy to be for this condition (see §5). It is clear that 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 $<$. They branch into the future but are linear in the past; each linear closed sub-order of a history is called a scenario. The sets of times in the history of each world are disjoint. "Times" could be instants or periods. A world and time are a "situation."

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 sub-relation 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; that 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., $a \in \mathbf{B}_{w,t}(x)$. Although, I will not pursue this 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., 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 can be similarly modeled; e.g., x always brakes before turning.

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 '$\text{seq}_i = \langle a_0, \ldots, a_{m-1}\rangle$' be a sequence of actions of member x_i. Then $[\![\text{seq}_i]\!] = \{\langle S, t, t'\rangle | (\exists t_0 \leq \ldots \leq t_m : t = t_0 \wedge t' = t_m \wedge (\forall j : j \in [1 \ldots m] \rightarrow \langle S, t_{j-1}, t_j\rangle \in [\![a_{j-1}]\!]_{x_i}))\}$; 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, $\text{seq} = \langle \text{seq}_1 \parallel \ldots \parallel \text{seq}_n\rangle$. The intension of 'seq' is the set of all subscenarios in the member sequences are executed starting together (and ending in any order). $[\![\text{seq}]\!] = \{\langle S, t, t'\rangle | (\exists e_0, \ldots e_{m-1} : \{e_0, \ldots e_{m-1}\} = \{0, \ldots, m-1\} \wedge t_{e_0} \leq \ldots \leq t_{e_{m-1}} \wedge (\forall j : j \in [1 \ldots m] \rightarrow \langle S, t, t_{e_j}\rangle \in [\![\text{seq}_j]\!]))\}$. $\langle S, t, t'\rangle \in [\![\text{seq}]\!]$ means that 'seq' starts at t and ends at t'; $[\![\text{seq}]\!] = \emptyset$ means that a subset of the member sequences is inconsistent. Sets of simultaneous actions can be treated as sets of sequences of length unity.

The formal language here is the predicate calculus, augmented with a temporal operator P (for "sometimes in the past"), and a two-place predicate, intends(agent, formula). The semantics is given relative to intensional models: it is standard for the predicate calculus; the predicate intends is considered below. Some auxiliary predicates are defined where needed. It is assumed throughout that operators for quoting and dequoting can be inserted where necessary.

4.2 Strategies

Following [25], I use "strategies" as abstract specifications of the behavior of an agent or a group. Strategies as defined here merely characterize an agent's behavior, possibly in quite coarse terms. There is no commitment to strategies being implemented as symbolic structures or as programs—they could just be the outcome of a particular architecture. Let Y be a strategy of group G; 'current(Y)' the part of Y now up for execution; and 'rest(Y)' the part of Y remaining after 'current(Y)' has been done. I 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 (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 (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 [19]), 'Y' is equivalent to 'current(Y); rest(Y).' Another obvious, but useful consequence of this is that 'current(Y)' is always of the form 'skip' or 'do(A)' or 'wait(A).'

But since 'skip' is the empty strategy, 'follows' 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. 'Follows' 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' follows 'do(A)' iff the group achieves A in doing it, and all subscenarios over which it is done preserve the background constraints. This allows agents to overact i.e., to do more than A requires (because of their habits, perhaps). 'Seq' follows 'wait(A)' iff it terminates just when A occurs, and all subscenarios over which it is done preserve the background constraints.

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

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

4.3 Group Structure

Following [24], strategies are used here to abstractly characterize the behavior of groups. We can be agnostic about whether they are hard-wired, or obtained through planning, or explicitly represented. A group strategy is treated as a set of strategies of its members. The structure of a group is defined by the interactions among its members at the levels of strategies and reactive actions. The interactions among the members of a group can be seen as objectively determining their respective "roles" in the group.

1. **Strategic Interactions:**

   Some of these abstract interactions among agents involve illocutionary acts [2]; e.g., assertions and commands. Others involve the establishment of various conditions in the world by some members' strategies that other members' strategies rely on. E.g., in a football team, the receivers run the patterns that the quarterback asks them to; some players are supposed to clear the path for their teammates, and whenever a defensive player identifies a move, he acts to block it.

2. **Reactive Interactions:**

   These are the interactions among the members as they execute their strategies, and may be implicit in the way in which a particular group acts. The interactions determine the joint "habits" of the group, as it were. E.g., a player may simply obstruct an opposing player trying to tackle his teammate, even though his strategy is to run forward himself; good players react to their opponents' moves by pushing as hard as possible without running into their teammates or committing fouls.

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, $x_i \in G$); and I_S and I_R are, respectively, the sets of *restrictions* on the strategic and reactive interactions among the x_i. 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 follows Y_i [24]. 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; some important categories are described in [24].

4.4 Performing Strategies

I now introduce a predicate, 'continuation' on strategies. Intuitively, if the 'current' part of a strategy is in progress, its continuation is the entire strategy; if the 'current' part of a strategy is over, its continuation is just the 'rest' of it. Define 'continuation(D, Y_G)' as the group strategy whose i_{th} element, for $i \in D$, is 'rest(Y_i),' and for $i \notin D$, 'Y_i.' Intuitively, D is the subset of members all of whom have *done* their 'current' parts. Then, a group performs a strategy along a particular sequence iff its members 'follow' the 'current' parts of their individual strategies while interacting in the right way and, as they complete the 'current' parts of their strategies, go on to do the next part.

In the following definitions, ϵ denotes the empty sequence, s' is a prefix of sequence s, and s'' the rest of it. The members of G follow the 'current' parts of their strategies as s' occurs, and the remaining part of the strategies as s'' occurs. Since the strategies need not proceed in lock-step, the restrictions are 'met' by all of s, rather than just s'.

$M \models_{w,t}$ performs$(x, \epsilon, \text{skip})$

$M \models_{w,t}$ performs(G, s, Y) iff $(\forall r : r \in (I_S \cup I_R) \rightarrow \text{meets}(r, s)) \wedge (\exists s', s'' : s' \neq \epsilon \wedge s = s' \circ s'' \wedge (\forall i : i \in [1, \ldots, n] \rightarrow M \models_{w,t} \text{follows}(x_i, s'_i, \text{current}(Y_i))) \wedge (\forall S : S \in \mathbf{S}_{w,t} \wedge (\exists t' : \langle S, t, t'\rangle \in [\![s']\!]) \rightarrow (\exists t' : \langle S, t, t'\rangle \in [\![s']\!] : (\exists t'', D : D \subseteq [1, \ldots, n] \wedge (\forall i : i \in D \rightarrow \langle S, t, t''\rangle \in [\![s'_i]\!]) \wedge (\forall j, t''' : j \in ([1, \ldots, n] \setminus D) \wedge \langle S, t, t'''\rangle \in [\![s'_j]\!] \rightarrow t'' < t''') \wedge M \models_{w,t''} \text{performs}(G, s'', \text{continuation}(D, Y_G))))))$

The recursive condition is applied every time a subset of the members (whose indices are in D) complete the 'current' part of their strategies: those members then move on to the next part of their strategies as the others simply carry on. The predicate, 'meets,' is meant to accommodate the restrictions on the interactions among agents, and has to be defined for each kind of restriction in $I_S \cup I_R$. Please see [24] for details.

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 sequences over which that group 'performs' that strategy.

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

This leaves the predicate 'comply-with' to still be defined. A strategy 'complies with' a restriction iff all sequences over which it is 'performed' 'meet' that restriction. Formally, this is quite simple:

$M \models_{w,t}$ comply-with(Y, r) iff $(\forall s :$ performs$(G, s, Y) \rightarrow$ meets$(s, r))$

This is included primarily to be able to express the requirement that an agent must explicitly represents some interactions with other agents, and explicitly intends for his strategies to satisfy them. In a way, such an agent participates consciously and willingly in his social relationships.

5 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 performance of the strategy it has then. Note that only the consequences of the successful performance 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.

- This definition considers as intentions only the *necessary* consequences of the performance of a 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.[1] 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.

- This definition considers the *performance* 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 performance of different strategies.

  Note also that it is not logical consequences, but rather the necessary consequences of performance that this definition uses. Thus the 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 externally all right 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

[1] It might have done that action *intentionally*, but need not have had the prior intention to pick it up—this distinction is described by Bratman [4, p. 119] and [5].

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

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

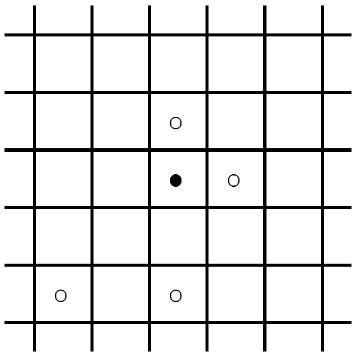
The predicate 'leads-to' is formally defined in §4.4; informally, it means that condition p is made true at some point on all scenarios where group G performs strategy Y; i.e., if G correctly performs Y, p would necessarily occur. The predicate 'has' captures the requirement that the strategy attributed 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 (§4.1):

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

5.1 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 DAI: the pursuit problem. This problem was introduced by Benda et al. in 1986 [3], but has been extensively studied by others since then [9, 27]. This problem has been analyzed from the perspective of discussing the impact and costs of different mechanisms of cooperation among the Blue agents, and from the perspective of the demands that declarative representations of this problem make. 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. The version here is taken from [27]. Briefly, the problem is as follows. We are given a finite two-dimensional grid of points (see Figure 1). 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 (if Red gets to the edge of the grid).



($\bullet$ = Red, $\circ$ = Blue)

Figure 1: The Pursuit Problem: an example configuration

Let the Blue agents be called B^1 through B^4, and Red be called R, and any of the five be called 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., 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 others would be analogous. This specification assumes that lower layers of the design are available to do the required actions. 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. Thus 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 that simply has the four substrategies. No restrictions have been imposed on this group. Thus the group intends that B^1 occupy the location above Red, B^2 the one to its right, and so on. In other words, the condition intended by the group as a whole is that it *win*. It is obvious from the construction that none of the Blue agents intends anything beyond occupying a certain location. Yet the group intends to achieve a win. This intention derives simply from the combinations of the members' intentions.

At the next level of detail, we might assign strategies to the Blue agents as follows. Let Y_1 be

while $\neg$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' is true when the given Blue agent abuts Red on the appropriate side for two clock cycles. 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 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 down before moving laterally. This could not have been concluded in the previous case. 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 where to go depending on the 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 control, and the other Blue agents its slaves. Now Y_1 could begin by first making an assignment of the locations surrounding Red to d-ifferent Blue agents; then *command* the other Blue agents to go there. Y_2, Y_3 and Y_4 could then simply be to any location about Red. The group would impose the strategic restriction that all commands from B^1 are obeyed [24]. 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.

5.2 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 in speaking of ant colonies. However, quite often it is desirable to ascribe an intention to a group only when its members participate more actively in it, or are aware to larger degrees of their role in the group and of the goal it is trying to achieve. This is required not just to give a truer account of several multiagent systems but also for certain practical reasons. A 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. Secondly, architectural considerations on the following of strategies (e.g., the perceptual hardware required) could be more easily described. E.g., a Blue agent who expected some information may keep a channel free; one who expected a neighboring agent to receive some important information could put off communicating with it. When the strategy of a group is itself only incompletely known, these requirements can be used to incrementally constrain the strategies of its members. In defining these stronger senses of group intentions, we can require that each member 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 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 $(\forall i : x_i \in G \rightarrow \ \text{has}(x_i, Y_i)) \wedge$

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

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, these components are not obvious from the strategies and structure of G. Since I am focusing on intentions in this paper, I simply define it to hold for all members of G.

- $M \models_{w,t} \text{believes}(G, \ \text{restriction}(G', r))$ iff $(\forall i : x_i \in G \rightarrow \ \text{believes}(x_i, \ \text{restriction}(G', r)))$

The use of the predicates introduced above, MB and 'believes,' makes this theory applicable to robust agents and groups. 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. While this task is not attempted in this paper, it is one that has been worked on by others [21, 23]. Briefly, these approaches relate the ascription of beliefs to the agent's actions; some further work is needed to make the connection with internal structure explicit.

While the addition of new constraints yields variations of this theory that appear closer to the traditional theories in terms of the logical axioms validated, this theory because of its consideration of the internal structure of groups provides a better motivation for those axioms.

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.[2] 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 intentions often have consequences that are not intended [4, 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. It is at this point that the power of the approach of [26] would be useful. That approach allows a semantically motivated definition of structural subsumption between the objects that intentions are defined over, but is too complex to be included here. Define intends$^i(G,p)$ as (here the superscript ^{i} is for "internal")

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

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)

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

For groups, 'intends-for' too may be defined in several ways, each sufficient for blocking consequential closure.

1. $M \models_{w,t} \text{intends-for}_1(G,Y,p)$ iff $(\exists i : x_i \in G \wedge \text{intends-for}_1(x_i,Y_i,p))$
2. $M \models_{w,t} \text{intends-for}_2(G,Y,p)$ iff $(\forall i : x_i \in G \wedge \text{intends-for}_2(x_i,Y_i,p))$

[2] 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).

3. $M \models_{w,t}$ intends-for$_3(G, Y, p)$ iff
   $(\forall i : x_i \in G \wedge$ MB$(x_1, \ldots, x_n;$ intends-for$_3(x_i, Y_i, p)))$

The definition of 'intends' with has$_4$ and intends-for$_3$ corresponds closely to that of Grosz and Sidner. While the members of a group are required to meet all the restrictions of that group, they may do so implicitly; i.e., without intending their strategies to actually contribute to the meeting of the restrictions imposed by the group. We can make this requirement explicit by adding the constraint that intends-for$(G, Y, p) \rightarrow (\forall i, r : r \in I_S \rightarrow$ intends-for$(x_i, Y_i,$ comply-with$(Y_i, r))$. This makes it explicit that the members' strategies are such that they try to meet all the restrictions. This models a group of perfectly cooperative agents who are completely introspective about their group.

6 Some Consequences

In this section, I enumerate some of the obvious formal consequences of the above definitions. For brevity, 'intends' is used for both intends^e and intends^i. Validity is indicated by '$\models$.'

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

   Valid when has$_1$ and has$_2$ are used in the definition of 'intends.' Unlike in the theories of Grosz and Sidner [13], and Cohen and Levesque [6], 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.

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

   P stands for "sometimes in the past." Valid for has$_1$ and has$_2$. If two members of a group have strategies that lead-to p and q, respectively, there is a group strategy that includes those strategies, and leads-to p and q in some arbitrary order. Note that this may hold vacuously because the appropriate group strategy may be impossible to perform successfully (perhaps due to the interactions required). Invalid for intends^i since it is possible that no member 'intends-for' the complex condition on the right hand side.

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

   Valid for has_1 and has_2. Valid since x's strategy is a part of the group's strategy, and intends^e allows consequential closure. Valid also for intends^i with intends-for_1, but not with intends-for_2 or intends-for_3.

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

   Valid also for intends^i with intends-for_1 or intends-for_2, but not with intends-for_3.

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

   Not valid since the group's intention may be due to a combination of its members' intentions; e.g., item 2 above.

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

   The group's intention may be a combination of its member's strategies; e.g., item 2 above. Valid, however, for intends^i with intends-for_3, as in the theory of Grosz and Sidner.

7 Conclusions

The theory presented here 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. The notion of strategies, and of their being followed reactively by agents and groups, provides a neat link to the theory of group ability. 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

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

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

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

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

[5] Michael E. Bratman. What is intention? Technical Report 69, Center for the Study of Language and Information, Stanford, CA, March 1987.

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

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

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

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

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

[11] Jacques Ferber. The framework of eco-problem solving. In *2nd European Workshop on the Modeling of Autonomous Agents in a Multi-Agent World*, August 1990.

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

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

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

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

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

[17] Steve Hickman and Martin Shiels. Situated action as a basis for cooperation. In *2nd European Workshop on the Modeling of Autonomous Agents in a Multi-Agent World*, August 1990.

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

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

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

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

[22] John R. Searle. Collective intentions and actions. In P. Cohen, J. Morgan, and M. Pollack, editors, *SDF Benchmark Series: Intentions in Communication*. MIT Press, Cambridge, MA, 1990.

[23] Nigel Seel. Formalising first-order intentional systems theory. Technical report, STC Technology Ltd., Harlow, Essex, UK, 1989.

[24] Munindar P. Singh. Group ability and structure. In *2nd European Workshop on the Modeling of Autonomous Agents in a Multi-Agent World*, August 1990.

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

[26] Munindar P. Singh and Nicholas M. Asher. Towards a formal theory of intentions. In *European Workshop on Logics in Artificial Intelligence*, September 1990.

[27] Larry M. Stephens and Matthais Merx. The effect of agent organization on the performance of DAI systems. *IEEE Transactions on Systems, Man and Cybernetics*, 1989. Submitted.