Teamwork*

Philip R. Cohen
Artificial Intelligence Center
and
Center for the Study of Language and Information
SRI International

and

Hector J. Levesque†
Dept. of Computer Science
University of Toronto

*This research was supported by a grant from the National Aeronautics and Space Administration to Stanford University (subcontracted to SRI International) for work on "Intelligent Communicating Agents," by a contract from ATR International to SRI International, and by a gift from the System Development Foundation. The second author was supported in part by a grant from the Natural Sciences and Engineering Research Council of Canada. This is a slightly revised version of a paper that appears in *Noûs 25/4*, 1991.

†Fellow of the Canadian Institute for Advanced Research.

1 Introduction

What is involved when a group of agents decide to do something together? Joint action by a team appears to involve more than just the union of simultaneous individual actions, even when those actions are coordinated. We would not say that there is any teamwork involved in ordinary automobile traffic, even though the drivers act simultaneously and are coordinated (one hopes) by the traffic signs and rules of the road. But when a group of drivers decide to do something together, such as driving somewhere as a convoy, it appears that the group acts more like a single agent with beliefs, goals, and intentions of its own, over and above the individual ones.

But given that actions are performed by individuals, and that it is individuals who ultimately have the beliefs and goals that engender action, what motivates agents to form teams and act together? In some cases, the answer is obviously the inherent value in doing something together, such as playing tennis, performing a duet, or dancing. These are examples of activities that simply cannot be performed alone. But in many cases, team activity is only one way among many of achieving the goals of the individuals. What benefits do agents expect to derive from their participation in a group effort?

In this paper, we attempt to provide an answer to these questions. In particular, we argue that a joint activity is one that is performed by individuals sharing certain specific mental properties. We show how these properties affect and are affected by properties of the participants. Regarding the benefits of teamwork, we show that in return for the overhead involved in participating in a joint activity, an agent expects to be able to share the load in achieving a goal in a way that is robust against certain possible failures and misunderstandings.

In the next section, we sketch our methodology and the adequacy criteria that have guided us. In section 3, we motivate certain aspects of our account by looking in detail at the example of a convoy. We then review the notion of an individual intention, and build on it for the joint case in sections 4 and 5. Then, in sections 7 and 8 , we discuss how our account satisfies the adequacy criteria we have laid out, and how this account relates to others. Finally, we draw some general conclusions.

2 Approach

The account of joint action presented here should probably not be regarded as a descriptive theory. We are primarily concerned with the design of artificial agents, under the assumption that these agents may need to interact with other agents (including people) having very different constitutions. At this stage, what we seek are reasonable *specifications,* that is, properties that a design should satisfy, and that would then lead to desirable behaviour. Thus we are not so much interested in characterizing some natural concept of joint activity; rather, we want to specify an idealized concept that has appropriate consequences. From this point of view, our discussion is in terms of what a specification guarantees, those properties an agent or group of agents satisfying the specification must have, as well as what the specification allows, those properties not ruled out by the specification. We attempt to guard against specifications that are too weak, in that they would fail to guarantee intuitively appropriate outcomes, as well as specifications that are too strong, in that they would place unreasonable demands on agents.

In our previous work [5], we have presented a belief-goal-commitment model of the mental states of individuals in which intentions are specified not as primitive mental features, but as internal

commitments to perform an action while in a certain mental state. Our notion of commitment, in turn, was specified as a goal that persists over time. A primary concern of the present research is to investigate in what ways a team is in fact similar to an aggregate agent, and to what extent our previous work on individual intention can be carried over to the joint case. Hence, we continue our earlier development and argue for a notion of *joint intention*, which is formulated as a joint commitment to perform a collective action while in a certain shared mental state, as the glue that binds team members together.

To achieve a degree of realism required for successful autonomous behaviour, we model individual agents as situated in a dynamic, multi-agent world, as possessing neither complete nor correct beliefs, as having changeable goals and fallible actions, and as subject to interruption from external events. Furthermore, we assume that the beliefs and goals of agents need not be known to other agents, and that even if agents start out in a state where certain beliefs or goals are shared, this situation can change as time passes.

This potential divergence of mental state clearly complicates our task. If we could limit ourselves to cases where every agent knew what the others were doing, for instance, by only considering joint actions that can be performed publicly, it would be much simpler to see how a collection of agents could behave as a single agent, because so much of their relevant beliefs would be shared.

On the other hand, it is precisely this potential divergence that makes joint activity so interesting: agents will not necessarily operate in lock step or always be mutually co-present, so there will be tension in trying to keep the team acting as a unit. Indeed, a primary goal of this research is to discover what would hold the team together, while still allowing the members to arrive at private beliefs about the status of the shared activity. In other words, even if we are willing to assume that everything is progressing smoothly during some shared activity, we will still be concerned with cases where, for example, one of the agents no longer has the belief that some other agent intends to do her share.

Moreover, it is this divergence among the agents that makes communication necessary. Whereas the model of individual intention in our earlier work [5, 6] was sufficient to show how communicative acts were defined in terms of beliefs and intentions, and could be used to achieve various goals, it did so only from the perspective of each individual agent, by constraining the rational balance that agents maintain among their own beliefs, goals, commitments, intentions, and actions. But special communicative demands are placed on agents involved in joint activities, and we wish to examine how these arise as a function of more general constraints on team behaviour.

Before looking at an example of the sort of joint activity we have in mind and possible specifications of the underlying team behaviour, we briefly list further questions that we expect our theory to address, in addition to those cited above:

Joint intentions leading to individual ones: As we said above, ultimately, it is agents that act based on their beliefs, goals, and intentions. How then do the joint beliefs, goals, and intentions of teams lead to those of the individuals, so that anything gets done? Typically, teams will be involved in joint activities that consist of many parts performed concurrently or in sequence. How do joint intentions to perform complex actions lead to appropriate intentions to perform the pieces? Assuming that an agent will only intend to do her own actions, what is her attitude towards the others' share?

The functional role of joint intentions: Bratman [2] has argued that in the case of individuals, intentions play certain functional roles: they pose problems for agents, which can be solved by

means-end analysis; they rule out the adoption of intentions that conflict with existing ones; they dispose agents to monitor their attempts to achieve them; and, barring major changes, they tend to persist. Which of these roles have analogues for teams?

Communication required: Any theory of joint action should indicate when communication is necessary. What do agents need to know (and when) about the overall activity, about their own part, and about the other agents' shares? Should agents communicate when the joint action is to begin, when one agent's turn is over, when the joint action is finished, when the joint action is no longer needed? How does communication facilitate the monitoring of joint intentions?

3 A Convoy Example

If team behaviour is more than coordinated individual behaviour, how does it work? This question is perhaps best answered by considering what would happen in the case of a convoy example without the right sort of joint intention.

Suppose we have two agents Alice and Bob; Bob wants to go home, but does not know his way, but knows both that Alice is going near there also and that she does know the way. Clearly Alice and Bob do not have to do anything together for Bob to get home; Bob need only follow Alice. In many circumstances, this plan would be quite adequate.

But it does have problems. For example, Alice might decide to drive very quickly through traffic, and Bob may be unable to keep up with her. It would be much better, from Bob's point of view, if Alice knew that he intended to follow her until he finds his way, counting on the fact that Alice, being a kind soul, would plan on keeping him in sight. Let us say that Bob arranges for Carl to tell her what he is going to do. Then, assuming she is helpfully disposed, she would not speed away. However, there is no reason for her to know that Bob is the one who sent Carl. As far as she is concerned, Bob might not know that she knows what he is up to. In particular, she would not expect Bob to signal her when he knows his way. So if, for example, Bob starts having car trouble and needs to pull over, Alice may very well speed off, believing that all is now well.

Realizing this is a possibility, Bob might try to get around it. He might get Carl also to tell Alice that he (Bob) is the one who asked Carl to talk to Alice. This would ensure that Alice was aware of the fact that Bob knew that she was being told. Assuming now that all goes well, at this point, both Bob and Alice would have appropriate intentions, both would know that the other had such intentions, and both would know that the other knew that they had such intentions.

However, there is still room for misunderstanding. Alice might say to herself: "Carl told me that Bob sent him to talk to me. So Bob now knows that I know what his driving plans are. But does he know that Carl mentioned that it was Bob who sent him? I think Carl just decided on the spot to say that, and so Bob doesn't realize that I know that it was him. So although Bob knows that I know his driving plans, he thinks I found out more or less accidentally, and so he thinks I won't expect him to signal me when he finds his way." Such reasoning might not happen, of course, but if it did, again Alice might speed off when Bob runs into car trouble. In fact, the situation is slightly worse than this, since even if this incorrect reasoning does not take place, Bob could still believe that it has, and not want to pull over for fear of being misunderstood.

This is clearly not the kind of robustness one expects from a convoy. The whole point of driving *together* is precisely to be able to deal better with problems that occur en route. The kind

of misunderstanding that is taking place here–and it could go on to deeper levels–is due to the fact that although both parties have the right intentions and the right beliefs about each other (at whatever level), they lack mutual belief of what they have agreed to. This suggests that Bob should approach Alice directly and get her to agree to the convoy, so that the agreement would be common knowledge between both parties.

Without being too precise about what exactly this means at this stage, we can nonetheless think of this as a rough first proposal for a concept of joint intention, that is, the property that will hold the group together in a shared activity. In other words, we expect agents to first form future-directed joint intentions to act, keep those joint intentions over time, and then jointly act.

> **Proposal 1.** x and y *jointly intend* to do some collective action iff it is mutually known between x and y that they each intend that the collective action occur, and it is mutually known that they each intend to do their share (as long as the other does theirs).

As we will discuss later in section 8, something very much like this has been proposed in the literature. As above, and assuming a tight connection between intention and commitment, it does indeed guarantee that the two agents commit to achieving the goal. Moreover, it is common knowledge that they are committed in this way and that neither party will change their mind about the desirability of the activity. In addition, we can assume that there are no hidden obstacles, in that if both parties did their share, then Bob would indeed get home. But even with these strong assumptions, the specification by itself is still too weak, once we allow for a divergence of mental states.

To see this, consider two appropriate reasons for dropping participation in the convoy: first, Bob could come to realize that he now knows his way, and so the intended action has successfully terminated; second, Alice could realize that she does not know where Bob lives after all, and so the intended action cannot be performed. We assume that in each case the agent in question has no choice but to give up the intention to act, terminating the convoy. The problem is that while Bob and Alice are driving together, Alice may come to believe that Bob now knows his way, that the convoy is over, and then speed off. Or Bob may come to believe that Alice does not know the way, that the convoy is over, and plan to get home some other way. As above, even if neither party comes to such an erroneous conclusion, they could suspect that something similar is happening with the other, and again the convoy would fall apart. Although both parties still have the right intentions and start with the right mutual knowledge, there is nothing to prevent this mutual knowledge from dissipating as doubt enters either agent about the private beliefs of the other regarding the status of the activity. But, these are potential troubles we would expect a joint activity to overcome.

More precisely, the problem with the first proposal is that although it guarantees goals and intentions that will persist suitably in time, it does not guarantee that the mutual knowledge of these goals and intentions will persist. So a second proposal is this:

> **Proposal 2.** x and y *jointly intend* to do some action iff it is mutually known between x and y that they each intend that the collective action occur, and also that they each intend to do their share as long as the other does likewise, and this mutual knowledge persists until it is mutually known that the activity is over (successful, unachievable, irrelevant).

This is certainly strong enough to rule out doubt-induced unraveling of the team effort, since both

parties will know exactly where they stand until they arrive at a mutual understanding that they are done.

The trouble with this specification is that, allowing for the divergence of mental states, it is too strong. To see this, suppose that at some point, Alice comes to realize privately that she does not know where Bob lives after all. The intention to lead Bob home is untenable at that point, and so there is no longer mutual belief that both parties are engaged in the activity. But to have been involved in a joint intention (in proposal 2) meant keeping that intention until it was mutually believed to be over. Since under these circumstances, it is not now mutually believed to be over, we are led to the counterintuitive conclusion that there was not really a joint intention to start with. The specification is too strong because it stipulates at the outset that the agents must mutually believe that they will each have their respective intentions until it is mutually known that they do not. It therefore does not allow for private beliefs that the activity has terminated successfully or is unachievable.

In section 5, we will propose more precisely a third specification for joint intention that lies between these two in strength and avoids the drawbacks of each. Roughly speaking we consider what one agent should be thinking about the other during the execution of some shared activity:

- The other agent is working on it (the normal case), or
- The other agent has discovered it to be over (for some good reason).

We then simply stipulate that for participation in a team, there is a certain *team overhead* to be expended, in that, in the second case above, it is not sufficient for a team member to come to this realization privately, she must make this fact mutually known to the team as a whole. As we will see, if we ensure that mutual knowledge of *this* condition persists, we do get desirable properties.

To see this in detail, we first briefly describe our analysis of individual commitment and intention, and then discuss the joint case.

4 Individual Commitment and Intention

Our formal account of individual and joint commitments and intentions [5, 15] is given in terms of beliefs, mutual beliefs, goals, and events. In this paper, we will not present the formal language, but simply describe its features in general terms. At the very lowest level, our account is formulated in a modal quantificational language with a possible-world semantics built out of the following primitive elements.

Events: We assume that possible worlds are temporally extended into the past and future, and that each such world consists of an infinite sequence of primitive events, each of which is of a *type* and can have an *agent*.[1]

Belief: We take belief to be what an agent is sure of, after competing opinions and wishful thinking are eliminated. This is formalized in terms of an accessibility relation over possible worlds in the usual way: the accessible worlds are those the agent has ruled capable of being the *actual* one. Beliefs are the propositions that are true in all these worlds. Although beliefs will normally change over time, we assume that agents correctly remember what their past beliefs were.

Goal: We have formalized the notion of goal also as accessibility over possible worlds, where the accessible worlds have become those the agent has selected as *most desirable*. Goals are the propositions that are true in all these worlds. As with belief, we presume that conflicts among choices and beliefs have been resolved. Thus, we assume that these chosen worlds are a subset of the belief-accessible ones, meaning that anything believed to be currently true must be chosen, since the agent must rationally accept what cannot be changed. However, one can have a belief that something is false now and a goal that it be true later, which is what we call an *achievement goal.* Finally, we assume agents always know what their goals are.

Mutual belief: The concept of mutual belief among members of a group will be taken to be the usual infinite conjunction of beliefs about other agents' beliefs about other agents' beliefs (and so on to any depth) about some proposition. Analogous to the individual case, we assume that groups of agents correctly remember what their past mutual beliefs were.

This account of the attitudes suffers from the usual possible-world problem of logical omniscience (see [14], for example), but we will ignore that difficulty here. Moreover, we will take *knowledge* simply (and simplistically) to be true belief, and *mutual knowledge* to be true mutual belief.

To talk about actions, we will build a language of *action expressions* inductively out of primitive events, and complex expressions created by action-forming operators for sequential, repetitive, concurrent, disjunctive, and contextual actions, where contextual actions are those executed when a given condition holds, or resulting in a given condition's holding. These dynamic logic primitives are sufficient to form a significant class of complex actions, such as the "if-then-else" and "while-loops" familiar from computer science [12]. In all cases, the agents of the action in question are taken to be the set of agents of any of the primitive events that constitute the performance of the action. To ground the earlier definition of collective action in the formal framework, we note that although a complex collective action may involve the performance by one agent of individual actions sequentially, repetitively, disjunctively, or concurrently with the performance of other individual actions by other agents, the collection of agents are not necessarily performing the action *together*, in the sense being explained in this paper.

For our purposes, it is not necessary to talk about actions with respect to arbitrary intervals (and thus have variables ranging over time points), but merely to have the ability to say that an action is happening, has just happened, and will happen next, with the implicit quantification that implies. It is also useful to define (linear) temporal expressions from these action expressions, such as a proposition's being *eventually*, *always*, or *never* true henceforth; similar expressions can be defined for the past. Finally, we say that a proposition remains true *until* another is true, with the obvious interpretation: if at some point in the future the former proposition is false, there must be an earlier future point where the latter is true.

4.1 Individual Commitment

Based on these primitives, we define a notion of individual commitment called persistent goal.[2]

Definition: An agent has a *persistent goal* relative to q to achieve p iff

1. she believes that p is currently false;

2. she wants p to be true eventually;
3. it is true (and she knows it) that (2) will continue to hold until she comes to believe either that p is true, or that it will never be true, or that q is false.

Some important points to observe about individual commitments are as follows: once adopted, an agent cannot drop them freely; the agent must keep the goal at least until certain conditions arise; moreover, other goals and commitments need to be consistent with them; and, agents will try again to achieve them should initial attempts fail. Clause 3 states that the agent will keep the goal, subject to the aforementioned conditions, in the face of errors and uncertainties that may arise from the time of adoption of the persistent goal to that of discharge.

Condition q is an irrelevance or "escape" clause, which we will frequently omit for brevity, against which the agent has relativized her persistent goal. Should the agent come to believe it is false, she can drop the goal. Frequently, the escape clause will encode the network of reasons why the agent has adopted the commitment. For example, with it we can turn a commitment into a subgoal, either of the agent's own supergoal, or of a (believed) goal of another agent. That is, an agent can have a persistent goal to achieve p relative to her having the goal of achieving something else. Note that q could in principle be quite vague, allowing disjunctions, quantifiers, and the like. Thus, we need not specify precisely the reasons for dropping a commitment. In particular, it could be possible to have a commitment to p relative to p being the most favored of a set of desires; when those rankings change, the commitment could be dropped. However, most observers would be reluctant to say that an agent is committed to p if the q in question is sufficiently broad, for example, such as that the agent could not think of anything better to do.

Finally, it is crucial to notice that an agent can be committed to another agent's acting. For example, an agent x can have a persistent goal to its being the case that some other agent y has just done some action. Just as with committing to her own actions, x would not adopt other goals inconsistent with y's doing the action, would monitor y's success, might request y to do it, or help y if need be. Although agents can commit to other's actions, they do not intend them, as we will see shortly.

4.2 Individual Intention

We adopt Bratman's [2] methodological concern for treating the future-directed properties of intention as primary, and the intention-in-action properties as secondary, contra Searle [20, 21]. By doing so, we avoid the notoriously difficult issue of how an intention self-referentially causes an agent to act, as discussed in [20], although many of those properties are captured by our account. Rather, we are concerned with how adopting an intention constrains the agents' adoption of other mental states.

An intention is defined to be a commitment to act in a certain mental state:

> **Definition:** An agent *intends* relative to some condition to do an action just in case she has a persistent goal (relative to that condition) of having done the action and, moreover, having done it, believing throughout that she is doing it.

Intentions inherit all the properties of commitments (e.g., tracking, consistency with beliefs and other goals) and also, because the agent knows she is executing the action, intention inherits properties that emerge from the interaction of belief and action. For example, if an agent intends

to perform a conditional action, for which the actions on the branches of the conditional are different, then one can show that, provided the intention is not dropped for reasons of impossibility or irrelevance, eventually the agent will have to come to a belief about the truth or falsity of the condition. In our earlier paper [5], we also show how this analysis of intention satisfies Bratman's [2, 3] functional roles for intentions and solves his "package deal" problem, by not requiring agents also to intend the known side-effects of their intended actions, despite our possible-world account of belief and goal.

Typically, an intention would arise within a subgoal-supergoal chain as a decision to do an action to achieve some effect. For example, here is one way to come to intend to do an action to achieve a goal. Initially the agent commits to p becoming true, without concern for who would achieve it or how it would be accomplished. This commitment is relative to q, so if the agent comes to believe q is false, she can abandon the commitment to p. Second, the agent commits to a or b as the way to achieve p, relative to the goal of p being true. Thus, she is committing to one means of achieving the goal that p be true. Third, the agent chooses one of the actions (say, a) and forms the intention to do it, that is, commits to doing a knowingly. The intention could be given up if the agent discovers that she has achieved p without realizing it, or if any other goal higher in the chain was achieved. For example, the intention might be given up if she learns that some other agent has done something to achieve q.[3] This example of intention formation illustrates the pivotal role of the relativization condition that structures the agent's network of commitments and intentions. We now turn to the joint case.

5 Joint Commitment

How should the definition of persistent goal and intention be generalized to the case where a group is supposed to act like a single agent? As we said earlier in the discussion of Proposal 2, joint commitment cannot be simply a version of individual commitment where a team is taken to be the agent, for the reason that the team members may diverge in their beliefs. If an agent comes to think a goal is impossible, then she must give up the goal, and fortunately knows enough to do so, since she believes it is impossible. But when a member of a team finds out a goal is impossible, the team as a whole must again give up the goal, but *the team does not necessarily know enough to do so.* Although there will no longer be mutual belief that the goal is achievable, there need not be mutual belief that it is *unachievable.* Moreover, we cannot simply stipulate that a goal can be dropped when there is no longer mutual belief, since that would allow agreements to be dissolved as soon as there was uncertainty about the state of the other team members. This is precisely the problem with the failed convoy discussed above. Rather, any team member who discovers privately that a goal is impossible (has been achieved, or is irrelevant) should be left with a goal to make this fact known to the team as a whole. We will specify that before this commitment can be discharged, the agents must in fact arrive at the mutual belief that a termination condition holds; this, in effect, is what introspection achieves in the individual case.

We therefore define the state of a team member nominally working on a goal as follows.

> **Definition:** An agent has a *weak achievement goal* relative to q and with respect to a team to bring about p if either of these conditions holds:
>
> - The agent has a normal achievement goal to bring about p, that is, the agent does not yet believe that p is true and has p eventually being true as a goal.

- The agent believes that p is true, will never be true, or is irrelevant (that is, q is false), *but* has as a goal that the status of p be mutually believed by all the team members.

So this form of weak goal involves four cases: either she has a real goal, or she thinks that p is true and wants to make that mutually believed,[4] or similarly for p never being true, or q being false.

A further possibility, that we deal with only in passing, is for an agent to discover that it is impossible to make the status of p known to the group as a whole, when for example, communication is impossible. For simplicity, we assume that it is always possible to attain mutual belief and that once an agent comes to think the goal is finished, she never changes her mind.[5] Among other things, this restricts joint persistent goals to conditions where there will eventually be agreement among the team members regarding its achievement or impossibility.[6]

The definition of joint persistent goal replaces the "mutual goal" in Proposal 2 by this weaker version:

> **Definition:** A team of agents have a *joint persistent goal* relative to q to achieve p just in case
>
> 1. they mutually believe that p is currently false;
> 2. they mutually know they all want p to eventually be true;
> 3. it is true (and mutual knowledge) that until they come to mutually believe either that p is true, that p will never be true, or that q is false, they will continue to mutually believe that they each have p as a weak achievement goal relative to q and with respect to the team.

Thus, if a team is jointly committed to achieving p, they mutually believed initially that they each have p as an achievement goal. However, as time passes, the team members cannot conclude about each other that they still have p as an achievement goal, but only that they have it as a *weak* achievement goal; each member allows that any other member may have discovered privately that the goal is finished (true, impossible, or irrelevant) and be in the process of making that known to the team as a whole. If at some point, it is no longer mutually believed that everyone still has the normal achievement goal, then the condition for a joint persistent goal no longer holds, even though a mutual belief in a weak achievement goal will continue to persist. This is as it should be: if some team member privately believes that p is impossible, even though the team members continue to share certain beliefs and goals, we would not want to say that the team is still committed to achieving p.

The first thing to observe about this definition is that it correctly generalizes the concept of individual persistent goal, in that it reduces to the individual case when there is a single agent involved.

> **Theorem:** If a team consists of a single member, then the team has a joint persistent goal iff that agent has an individual persistent goal.

The proof is that if an agent has a weak goal that persists until she believes it to be true or impossible, she must also have an ordinary goal that persists.

It can also be shown that this definition of joint commitment implies individual commitments from the team members.

Theorem: If a team has a joint persistent goal to achieve p, then each member has p as an individual persistent goal.

To see why an individual must have p as a persistent goal, imagine that at some point in the future the agent does not believe that p is true or impossible to achieve. Then there is no mutual belief among the whole team either that p is true or that p is impossible, and so p must still be a weak goal. But under these circumstances, it must still be a normal goal for the agent. Consequently, p persists as a goal until the agent believes it to be satisfied or impossible to achieve. A similar argument also shows that if a team is jointly committed to p, then any *subteam* is also jointly committed. This generalization will also apply to other theorems about intention presented below.

So if agents form a joint commitment, they are each individually committed to the same proposition p (relative to the same escape condition q). If p is the proposition that the agents in question have done some collective action constructed with the action-formation operators discussed above, then each is committed to the entire action's being done, including the others' individual actions that comprise the collective. Thus, one can immediately conclude that agents will take care to not foil each other's actions, to track their success, and to help each other if required.

Furthermore, according to this definition, if there is a joint commitment, agents can count on the commitment of the other members, first to the goal in question and then, if necessary, to the mutual belief of the status of the goal. This property is captured by the following theorem, taken from our earlier work [15].

Theorem: If a team is jointly committed to some goal, then under certain conditions, until the team as a whole is finished, if one of the members comes to believe that the goal is finished but that this is not yet mutually known, she will be left with a *persistent* goal to make the status of the goal mutually known.

In other words, once a team is committed to some goal, then any team member that comes to believe privately that the goal is finished is left with a a *commitment* to make that fact known to the whole team. So, in normal circumstances,[7] a joint persistent goal to achieve some condition will lead to a private commitment to make something mutually believed. Thus, although joint persistent goal was defined only in terms of a weak goal—a concept that does not by itself incorporate a commitment—a persistent goal does indeed follow.

This acquisition of a commitment to attain mutual belief can be thought of as the team overhead that accompanies a joint persistent goal. A very important consequence is that it predicts that *communication* will take place, as this is typically how mutual belief is attained, unless there is co-presence during the activity. Thus, at a minimum, the team members will need to engage in communicative acts to attain mutual belief that a shared goal has been achieved.

6 Joint Intention

Just as individual intention is defined to be a commitment to having done an action knowingly, joint intention is defined to be a joint commitment to the agents' having done a collective action, with the agents of the primitive events as the team members in question, and with the team acting in a joint mental state.

Definition (and Proposal 3): A team of agents *jointly intends,* relative to some escape condition, to do an action iff the members have a joint persistent goal relative to

that condition of their having done the action and, moreover, having done it mutually believing throughout that they were doing it.[8]

That is, the agents are jointly committed to its being the case that throughout the doing of the action, the agents mutually believe they are doing it.

Next, we examine some of the important properties of joint intention.

6.1 Properties of Joint Intention

Given that joint intention is a property of a group of agents, but that only individual agents act, what is the origin of the individual intentions that lead those agents to perform their share? We have shown that joint persistent goals imply individual goals among the team members. We now wish to show a similar property for joint intentions.

First, observe that joint intention implies individual intention when one agent is the only actor.

> **Theorem:** If a team jointly intends to do an action, and one member believes that she is the only agent of that action, then she privately intends to do the action.

This holds because joint commitment entails individual commitment, and mutual belief entails individual belief. Of importance is the added condition that the agent must believe herself to be the only agent of the action. As desired, we do not allow agents to intend to perform other agents' actions, although they can be committed to them.

In the case of multi-agent actions, we will only consider two types: those that arise from more basic actions performed *concurrently*, and those that are formed from a *sequence* of more basic actions.

6.2 Jointly Intending Concurrent Actions

Consider the case of two agents pushing or lifting a heavy object, or one bracing an object while the other acts upon it. First, we need the following property of individual intention.

> **Theorem:** An individual who intends to perform actions a and b concurrently intends to perform a (resp. b) relative to the broader intention.

The proof of this depends on the treatment of concurrency as a conjunction of actions performed over the same time interval. Hence, the conjuncts can be detached and treated separately. Note that the intention to perform a is only relative to the intention to do both parts together; should the agent come to believe that it is impossible to do b at all, she may very well not want to do a alone.

Analogously, for joint intention, the following holds.

> **Theorem:** If a team jointly intends to do a complex action consisting of the team members concurrently doing individual actions, then the individuals will privately intend to do their share *relative to the joint intention.*

In other words, agents who jointly intend concurrent actions also individually intend to do their parts as long as the joint intention is still operative. The proof of this parallels the proof that joint intention leads to individual intention in the case of single-agent actions. Individual intentions thus

persist at least as long as the joint intention does. But the commitment can be dropped if a team member discovers, for example, that some other team member cannot do her share.

Thus, an unrestricted individual intention, that is, an intention that is not relative to the larger intention, does *not* follow from a joint intention. Still, as with any joint persistent goal, even if one agent discovers privately that the joint intention is terminated, there will remain residual commitments to attain mutual belief of the termination conditions.

Notice also that agents are supposed to mutually believe, throughout the concurrent action, that they are performing it together. Thus, while the agents are performing their individual actions, they also each believe that together they are performing the group action.

6.3 Jointly Intending Sequential Actions

Next, we need to ascertain how joint intentions for sequential actions result in the agents acquiring their own individual intentions. This case is more complex, since temporal properties and execution strategies need to be considered.

6.3.1 Stepwise Execution

Consider, first, individual intention and action. Processors for programming languages in computer science are usually designed to step through a program "deliberately," by keeping track of what part of the action is being executed and, if there are conditions (such as for *if-then-else* actions), by ascertaining the truth or falsity of those conditions before proceeding with the computation or execution.

However, the framework we have adopted allows individual agents to be considerably more flexible in executing action expressions. For example, though an agent may know she is executing a repetitive or sequential action, she need not know where she is in the sequence. For example, an agent can click on a phone receiver a number of times and know that one of those clicks disconnects the line and produces a dial tone without ever having to know which click was the one that did it. Similarly, an agent need not know the truth value of a condition on an if-then-else action if the two branches share an initial sequence of events. So, for instance, to execute an action expressed as "if it is raining, then bring all your rain gear, otherwise just bring the umbrella" it is sufficient to get an umbrella before checking the weather, since that is required in either case. Only at the point at which those execution paths diverge will it be necessary for the agent to have a belief about the truth of the past condition.

This freedom may seem like unnecessary generality, although, as we will see, it plays an important role in the case of joint activity. However, one consequence it has is that an agent who intends to do a sequential action does *not* necessarily intend to perform the first step in the sequence, even relative to the larger intention. It is consistent with our specification that an agent can intend to do the sequence without expecting to know when the first part of that sequence is over. Thus, the reasons for dropping the commitments entailed in having an intention would not be present. Moreover, the agent need not intend to do the remainder of the sequence either: since she might not know when the first part has been completed, she might not know she is doing the second part. In other words, because one may not know when subactions start and stop, it is possible to execute a sequence of actions knowingly without knowingly executing the individual steps.

However, it is possible to stipulate a condition on the execution of a sequence that would guarantee the distribution of intention throughout the sequence: we can require the agent to

believe after each step both that the step was just done and that she is doing the remainder. We call this *stepwise execution*. That is, in the stepwise execution of a sequence, each step becomes a contextual action: it must be performed in a context where the agent has certain beliefs. In effect, this forces the agent to execute the action like a traditional programming language processor and leads to the following theorem.

> **Theorem:** If an agent intends to do a sequential action in a stepwise fashion, the agent also intends to do each of the steps, relative to the larger intention.

The proof of this theorem is that if an agent believes she has done the entire sequence in a stepwise fashion, she must believe that she had the belief at the relevant times about having done each step, and by the memory assumption of section 4, these beliefs cannot simply be after-the-fact reconstructions.[9]

6.3.2 Joint Stepwise Execution

Given that to obtain a seemingly desirable property of intending sequences which most agent designers implicitly assume, the agent must explicitly intend to execute the sequence in stepwise fashion, the freedom offered in our formulation of sequential action execution may seem like a dubious advantage. However, it has considerable merit when one considers joint action. Recall that one of our principles has been to maximize the similarity of joint commitments and intentions to their individual counterparts. If stepwise execution of actions were the *only* way to execute actions, and if, following the similarity principle, we applied that strategy to a team, we would thereby enforce a *joint* stepwise execution strategy, requiring the attainment of mutual belief after each step that the step had been accomplished and that the agents were embarking on the remainder.

But we do not want to *require* a team to always execute complex actions in lock step. There are many types of joint actions where such team overhead would be undesirable. Consider, for example, an expert and an apprentice performing a sequence together, where the expert has to do something immediately after the apprentice. The two may separate from one another, with only the expert being able to sense the apprentice, and hence know when it is her turn to act. In fact, it is possible that only the expert will know when the apprentice has successfully done his part. She may be free to continue the sequence, and report success of the whole enterprise, without reporting intermediate results. Thus, we want to allow for individuals to contribute privately to a sequence, when that is compatible with the performance of the overall activity. So, to allow for such actions, the joint intention to do a sequence must not require the agents to come to a mutual belief that each step has just been done successfully.

How, then, do team members get individual intentions in such cases? Essentially, all that is needed is that each agent know that it is her turn and what she is doing.

> **Theorem:** If a team jointly intends to do a sequential action, then the agent of any part will intend to do that part relative to the larger intention, provided that she will always know when the antecedent part is over, when she is doing her share, and when she has done it.

As always, the individual intentions are formed relative to the larger joint intention.

However, many joint activities, such as games and dialogue, *are* supposed to be performed in a joint stepwise fashion. For example, agents who jointly intend to play a set of tennis jointly intend

to play the first point. After the point, the agents must agree that it is over (and who won it) before proceeding. So, we need to allow for *both* forms of joint execution of a sequential action. Fortunately, our earlier analysis of individual action provides just the right kind of generalization and offers an immediate analogue for the joint case.

> **Theorem:** If a team intends to do a sequence of actions in a joint stepwise fashion, the agents of any of the steps will jointly intend to do the step relative to the larger intention.

As before, appropriate individual intentions and commitments will then follow from the joint intentions.

7 Meeting the Adequacy Criteria

In characterizing joint commitments and intentions, we have specified a notion of weak achievement goal as the property that persists and holds the group together. Given this, we have addressed our first adequacy criterion by showing the conditions under which joint intentions to perform simple actions, concurrent actions, and sequential actions entail the team members forming the relevant individual intentions. Joint intentions embody a precise notion of commitment, yet are not defined in terms of the individual intentions. Instead, both are defined in terms of the same primitives, and the individual intentions follow from the joint ones.

As we have seen, joint commitments give rise to individual commitments *relative to* the overarching joint commitment. Thus, the individual commitments are subsidiary to the joint ones and can be abandoned if the joint ones are given up. Moreover, because a jointly intended action requires the agents to mutually believe they are acting together, an agent does not merely believe she is acting alone. Rather, the agents believe their actions are part of and depend on the group's commitment and efforts.

Turning now to the functional role of joint intentions, our discussion of execution strategies implies that the adoption of a joint intention need not always lead to a process of joint problem-solving that culminates in a mutual belief among the team members regarding what each is to do. Rather, this property would only hold if it were necessary for the execution of the action or if the agents agreed to perform their actions in a more deliberate way, such as in a joint stepwise fashion.

However, joint intentions do form a "screen of admissibility" [2], analogous to those of individual commitments, because joint commitments and, hence, intentions must be consistent with the individual commitments. Just as agents cannot knowingly and intentionally act to make their individual commitments and intentions impossible to achieve, they similarly cannot act to make their joint commitments and joint intentions impossible. In particular, they will not knowingly and intentionally act to foil their team members' actions. On the other hand, if it is mutually known that one team member requires the assistance of another, our account predicts that the other will intend to help. All of these properties follow immediately from the fact that joint commitments entail individual commitments, and that these must be mutually consistent. For more specific analyses of the kinds of consistency predicted by our analysis of individual commitment and intention, see our more comprehensive paper [5].

In addition, as in the individual case, a group will monitor the success or failure of the joint effort, and, in particular, with joint stepwise execution, it will monitor the intermediate results as well. These results follow from the facts that agents who have jointly intended to do some collective

action are jointly committed to mutually believing that they are performing the action, and that they must ultimately come to mutually believe that they have done it or that it is impossible or is irrelevant.

As for the communication criterion, by showing that agents who have adopted a joint intention commit themselves to attaining mutual belief about the status of that intention, we derive commitments that may lead to communication. For example, in other of our papers on dialogues about a task [7, 8], we have analyzed how joint intentions to engage in the task lead to the discourse goals that underlie various speech acts.

8 Comparison with Other Analyses of Joint Intention

Numerous analyses of concepts similar to joint intention have been given. Tuomela and Miller [22] propose a conceptual analysis of an individual agent's "we-intending" a group action. Essentially, that agent must intend to do her part of the action and believe it is mutually believed that the other members of the team will do their parts as well. Power [18] is perhaps the earliest researcher within the artificial intelligence research community to be concerned with modeling joint activity. He defines a mutual intention to be each agent's having the intention to do her part and there being a mutual assumption that each agent has such intentions. Grosz and Sidner [10] propose a concept of shared plans, using Pollack's [17] analysis of plans and Goldman's [9] analysis of action. In their model, two agents have a shared plan if those agents mutually know that each agent intends to do her own part to achieve the jointly done action, and that each agent will do her part if and only if the other agent does likewise.

Though differing in detail, these analyses share a number of disadvantages as compared to the analysis proposed here. First, they do not make clear how, if at all, the agents are committed to a joint activity or to its parts, although Grosz and Sidner's come closest with their use of the biconditional relating agents' intentions. Specifically, they do not show how one agent can be committed to the other's acting, without stating that the agent intends the other agent's actions, an expression that would be ill-formed in most analyses of intention. Such commitment to the others' actions are important, since they would lead one agent to help another, to stay out of her way, etc., as we have described.

Second, even granting some notion of commitment inherent in their uses of the term 'intention', these analyses all possess the defects of Proposal 1: though the agents' intentions to do their parts may persist, there is no constraint on the persistence of the agents' *mutual beliefs* about those intentions. Hence, such analyses are dissolved by doubt. Finally, because there is no requirement to start or terminate joint actions with mutual belief, these analyses make no predictions for communication.

Searle [21] provides a different argument against approaches such as these, claiming that collective intentions are not reducible to individual intentions, even when supplemented with mutual beliefs. He claims to provide a counterexample of a group of students who have been jointly educated to be selfish capitalists, and mutually know their classmates have been similarly indoctrinated to compete vigorously, with the collective goal of serving humanity and themselves. The students are claimed to satisfy Tuomela and Miller's definition (and, by extension, Power's), but are not acting collectively.[10] On the other hand, Searle argues that had the students made a pact on graduation day to compete vigorously, their subsequent actions would constitute a joint activity.

Instead of reducing collective intention to some combination of individual intention and mutual

belief, Searle proposes a *primitive* construct, one not defined in terms of other concepts, for "we-intending" in which individual agents we-intend to do an action by means of the individual agents doing their share. By using a new primitive construct, Searle attempts to solve the problem addressed earlier, namely, how a group's collective intention leads the individual agents to form their own intentions.[11] Rather than propose a primitive construct for collective intention, we have shown that we can derive reasonable entailments and meet a substantial set of adequacy criteria by defining both joint and individual intentions in terms of the same set of primitive elements.

A major concern of the present paper has been to characterize joint commitment suitably so that it keeps a group together long enough to take action. Thus, it is crucial to our understanding that joint intention be regarded as a future-directed joint commitment. Although Searle's examples are motivated by cases of future-directed collective intention, Searle's analysis extends only his notion of intention-in-action [20] to the collective case. Thus, the analysis is silent about how a group could plan to do some action in the future, and about how such collective future-directed intentions could eventually result in the formation of a collective present-directed intentions.

9 Conclusions

At this point, we have exhibited some of the consequences of a group's adopting joint commitments and intentions. Once adopted, agents should be able to build other forms of interaction upon them. Here, we only have space to remark in passing on how this might work by looking briefly at contracts and agreements, speech acts, and dialogue.

First, an interesting extension of our analysis would be to describe how the properties that result from the adoption of joint commitments and intentions compare with those inherent in the formulation of valid contracts [1]. We suspect that informal versions of many of the properties of contracts can be found in our notion of a joint commitment, especially in cases where there can be disagreement about the final contractual outcome. Historically, contracts (in British contract law) were regarded as formalized agreements. Hence, if our account were to bear some relation to contract law, it would be through some understanding of what constitutes an agreement.

The clearest cases of joint activity are ones in which either an explicit or implicit agreement to act is operative, by which we mean some sort of mental state that agents enter into and the speech acts by which they do so. Although there surely is a complex interrelationship between having a joint intention and there being such an agreement in force, we have taken the concept of joint intention simply to be present in all agreements. For the purposes of this paper, the two concepts have been treated as one.

Future work will examine how speech acts of various kinds might be used to create agreements and, hence, joint commitments. Currently, our theory of speech acts [6] argues that the intended effect of a request is to get the addressee to form an individual commitment to do the requested action relative to the speaker's desire that he do so. Although the addressee may be individually committed, nothing in our account prevents the speaker from changing her mind, not notifying the addressee, and then deliberately making the requested action impossible. This would be a clear violation of tacit rules of social behaviour, but nothing in an individualistic account of the commitments entailed by acceding to a request would prevent it. The question remains, for designing artificial agents, should we augment the semantics of their speech acts to somehow make mutual promises or requests followed by acknowledgments yield joint commitments? And if not, where do the joint commitments come from?

One can also now imagine developing a more general account of dialogue, in which a theorist formally analyses the *social contract* implicit in dialogue in terms of the conversants' jointly intending to make themselves understood and to understand the other. From our perspective, the signals of understanding and requests for them, which are so pervasive in ongoing discourse [4, 16, 19], would thus be predictable as the means to attain the states of mutual belief that discharge this joint intention [8, 7]. More generally, if such an account of dialogue were successful, it might then be possible to formalize cooperative conversation in a way that leads to the derivation of Gricean maxims.

Finally, let us return to one of our original motivations, designing agents that can work together in groups. Research in artificial intelligence has in the main concentrated on the design of individual agents. If that work is successful (a big "if" indeed), there will undoubtedly be many agents constructed and let loose on the world. Without consideration of how they will cooperate and communicate with other agents, perhaps of dissimilar design, and with people, we risk a kind of "sorcerer's apprentice" scenario—once let loose, they cannot be controlled, and will compete with the other agents for resources in achieving their selfish aims. Joint commitments, we claim, can form the basis for a social order of agents, specifying how groups remain together in the face of unexpected events and the fallible and changeable nature of the agents' attitudes. If built according to our specifications, once such agents agree to cooperate, they will do their best to follow through.

Acknowledgments

Many thanks go to Michael Bratman, David Israel, Henry Kautz, Kurt Konolige, Joe Nunes, Sharon Oviatt, Martha Pollack, William Rapaport, Yoav Shoham, and Bonnie Webber for their valuable comments. The second author also wishes to acknowledge the Center for the Study of Language and Information and the Department of Computer Science of Stanford University, where he was a visitor during the preparation of this paper.

Notes

[1] Currently, we picture these events as occurring in a discrete synchronized way, but there is no reason not to generalize the notion to a continuous asynchronous mode, modeled perhaps by a function from the real numbers to the set of event types occurring at that point.

[2] This definition differs slightly from that presented in our earlier work [5], but that difference is immaterial here.

[3] Of course, the agent may still intend to achieve p again if she is committed to doing so *herself*.

[4] More accurately, we should say here that her goal is making it mutually believed that p *had been* true, in case p can become false again.

[5] For readers familiar with the results in distributed systems theory [11] in which it is shown that mutual *knowledge* is impossible to obtain for computers by simply passing messages, we point out that those results do not hold for mutual beliefs acquired by default, nor for agents that can be co-present or communicate instantly.

[6] Actually, agents do have the option of using the escape clause q to get around this difficulty. For example, $\neg q$ could say that there was an unresolvable disagreement of some sort, or just claim that an expiry date had been reached, or that the agents each no longer wants to have the joint intention. In such cases, mutual belief in $\neg q$ amounts to an agreement to dissolve the commitment regardless of the status of p.

[7] The normality conditions referred to here are merely that once the agent comes to a belief about the final status of the goal, she does not change her mind before arriving at a mutual belief with the others.

[8] A more precise version of this definition [15] also requires that they mutually know when they started.

[9] Another way to obtain a similar result might be to change the definition of persistent goal to say that an agent can drop her goal that p if she comes to believe that p *has been made true*, rather than is currently true. However, this introduces additional complexity, since one must be careful not to consider times when p was true before the adoption of the goal.

[10] Whether Searle's example also counters Grosz and Sidner's analysis, as claimed by Hobbs [13], is arguable. They may escape the example's force because of the biconditional in their definition: there must be mutual belief that each agent intends to do his part *iff* the other agent does likewise.

[11] In Tuomela and Miller's, Power's, and Grosz and Sidner's analyses, the means by which we-intentions, mutual intentions, and shared plans, respectively, lead agents to have individual intentions is no mystery: they are simply defined in terms of individual intention.

References

[1] P. S. Atiyah. *An Introduction to the Law of Contract.* Oxford University Press, Oxford, U. K., 1989.

[2] M. Bratman. *Intentions, Plans, and Practical Reason.* Harvard University Press, 1987.

[3] M. Bratman. What is intention? In P. R. Cohen, J. Morgan, and M. E. Pollack, editors, *Intentions in Communication.* MIT Press, Cambridge, Massachusetts, 1990.

[4] H. H. Clark and D. Wilkes-Gibbs. Referring as a collaborative process. *Cognition*, 22:1–39, 1986.

[5] P. R. Cohen and H. J. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42(3), 1990.

[6] P. R. Cohen and H. J. Levesque. Rational interaction as the basis for communication. In P. R. Cohen, J. Morgan, and M. E. Pollack, editors, *Intentions in Communication.* MIT Press, Cambridge, Massachusetts, 1990.

[7] P. R. Cohen and H. J. Levesque. Confirmations and joint action. In *Proceedings of the 12th International Joint Conference on Artificial Intelligence*, Sydney, Australia, August 1991. Morgan Kaufmann Publishers, Inc.

[8] P. R. Cohen, H. J. Levesque, J. Nunes, and S. L. Oviatt. Task-oriented dialogue as a consequence of joint activity. In *Proceedings of the Pacific Rim International Conference on Artificial Intelligence*, Nagoya, Japan, November 1990.

[9] A. I. Goldman. *A Theory of Human Action.* Princeton University Press, Princeton, New Jersey, 1970.

[10] B. Grosz and C. Sidner. Plans for discourse. In P. R. Cohen, J. Morgan, and M. E. Pollack, editors, *Intentions in Communication.* MIT Press, Cambridge, Massachusetts, 1990.

[11] J. Y. Halpern and Y. O. Moses. Knowledge and common knowledge in a distributed environment. In *Proceedings of the 3rd ACM Conference on Principles of Distributed Computing*, New York City, New York, 1984. Association for Computing Machinery.

[12] D. Harel. *First-Order Dynamic Logic.* Springer-Verlag, New York City, New York, 1979.

[13] J. R. Hobbs. Artificial intelligence and collective intentionality: Comments on Searle and on Grosz and Sidner. In P. R. Cohen, J. Morgan, and M. E. Pollack, editors, *Intentions in Communication.* MIT Press, Cambridge, Massachusetts, 1990.

[14] H. J. Levesque. A logic of implicit and explicit belief. In *Proceedings of the National Conference of the American Association for Artificial Intelligence*, Austin, Texas, 1984.

[15] H. J. Levesque, P. R. Cohen, and J. Nunes. On acting together. In *Proceedings of AAAI-90*, San Mateo, California, July 1990. Morgan Kaufmann Publishers, Inc.

[16] S. L. Oviatt and P. R. Cohen. Discourse structure and performance efficiency in interactive and noninteractive spoken modalities. *Computer Speech and Language*, 1991, in press.

[17] M. E. Pollack. Plans as complex mental attitudes. In P. R. Cohen, J. Morgan, and M. E. Pollack, editors, *Intentions in Communication*. M.I.T. Press, Cambridge, Massachusetts, 1990.

[18] R. Power. Mutual intention. *Journal for the Theory of Social Behavior*, 14(1):85–102, March 1984.

[19] E. A. Schegloff. Discourse as an interactional achievement: Some uses of unh-huh and other things that come between sentences. In D. Tannen, editor, *Analyzing discourse: Text and talk*. Georgetown University Roundtable on Languages and Linguistics, Georgetown University Press, Washington, D.C., 1981.

[20] J. R. Searle. *Intentionality: An Essay in the Philosophy of Mind*. Cambridge University Press, New York, New York, 1983.

[21] J. R. Searle. Collective intentionality. In P. R. Cohen, J. Morgan, and M. E. Pollack, editors, *Intentions in Communication*. M.I.T. Press, Cambridge, Massachusetts, 1990.

[22] R. Tuomela and K. Miller. We-intentions. *Philosophical Studies*, 53:367–389, 1988.