On Team Formation

Philip R. Cohen[1], Hector J. Levesque[2] and Ira Smith[1]

1 Introduction

The concept of joint action is at the core of numerous diverse research topics, including philosophical explorations of social action, studies of human dialogue, human-computer interaction, computer-supported collaborative work, multiagent systems, distributed artificial intelligence, distributed simulation, and contract law. It is therefore remarkable that so central a concept has received so little detailed analysis, in comparison with studies of individuals. However, in recent years, the study of joint action has begun to undergo more intense scrutiny, primarily from philosophers and researchers in artificial intelligence. These two disciplines often address similar topics but with different motivations, methodologies, tools, and criteria for success. Although this paper is inspired by philosophical work, it is squarely motivated by the concerns of building intelligent systems that are capable of collaborative behavior, either with a user, or with other such systems. Still, we hope that the paper sheds light on philosophical issues, and treats the subject of joint action at a sufficiently precise level to be illuminating of problems that any philosophical account needs to confront.

An important consequence of focusing on joint actions, rather than solely on individual actions, is the opportunity to rethink related theories. In particular, we claim that speech act theory will need to be recast in light of joint action theory since many of the basic illocutionary acts (e.g., requests, promises) are intimately involved in eatablishing, monitoring, and discharging joint activities. However, despite this tight relationship, no existing speech act theory provides guidance on this connection. This paper takes a first step in the direction of linking speech act theory and joint action theory by showing how various speech acts can be used to form and disband teams.

It is by now commonplace to observe that joint action is different from a collection of individual actions, even if they are coordinated. Agents can be acting in a coordinated fashion, as in ordinary automobile traffic, but not be acting together. Conversely, agents can be acting together, but not be coordinated except at the start and end of their joint action (e.g, see [36]) The key property distinguishing joint or collaborative action from mere coordinated action is the joint mental state of the participants.

The best way to explore what this mental state must be is to imagine a joint action going astray. Our favorite example is driving in a convoy, versus ordinary traffic. If one driver observes another leaving the road, there is no requirement that the observer take any specific action (except, perhaps, avoidance). However, if the drivers are in a convoy, and one driver is observed pulling off the road, the other drivers will do likewise. In ordinary traffic, it is no cause for concern if one driver loses sight of another. Moreover, there could be no serious consideration about whether that driver has changed his mind about his route or destination. In a convoy, however, losing sight of

[1]Center for Human-Computer Communication, Oregon Graduate Institute of Science and Technology, Beaverton, Oregon

[2]Dept. of Computer Science, University of Toronto, Toronto, Canada

one's fellow travellers, or doubting whether the leader is still leading *would* be causes for concern. Nonetheless, a convoy will continue, even should such doubts occur — joint actions do not simply fall apart when difficulties arise, but rather are robust to mistakes, failures, doubt, etc.

To explain joint actions, one must explain the cohesive force that binds team members together. Because this force is so strong, joint actions are not lightly undertaken. By teaming with a partner, an individual plans to share the load, to have mutual goals, to help when necessary, etc. Team members commit resources to their teammates, passing up opportunities inconsistent with the team's success. In support of these mutual commitments, teams involve an inherent overhead — in establishing, monitoring, and disbanding the team. A theory of joint action therefore needs to explain how this occurs.

Unfortunately, although there are numerous analyses of joint action in the literature (e.g, [4, 12, 19, 18, 44, 43, 36] how agents in fact form and disband teams is rarely discussed. In some cases, it is suggested that joint activities are formed by "agreements," but that concept is left relatively unformalized. To be of theoretical and practical utility, the informal notion of an agreement needs to be cashed out in terms of a precisely stated speech act theory. Once that is done, it should be possible to prove that a sequence of communicative actions in fact establishes a joint activity, and another sequence in fact discharges it. The contribution of this paper is to provide such an analysis.

The structure of the paper is as follows: First, we review the foundational analysis of individual commitment and intention, out of whose primitives are constructed the theory of joint commitment and joint intention. Then, those concepts are used to form definitions of communicative actions. In particular, we redefine directive actions such that they support team formation. Finally, we then show how these communicative actions can be used to form and disband teams.

2 Joint actions — collective actions by agents who share a joint intention

In our previous work [8], 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 [8, 10] 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 any 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, conditionally, 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?

The essence of team behavior is cooperation; if a team has a goal and a plan to achieve the goal, team members are not only individually committed to the success of their portion of the plan, they are also committed to the success of the overall team goal, and to the success of other team members achieving their portions of the overall joint plan. This explains why we would expect the other drivers in a convoy to stop to aid the disabled car. Moreover, everyone in the team knows and expects the other agents to act in this way. This joint commitment to each others' success, and to the overall success of the project is a major difference between a team and a group of subcontractors.

Without being too precise about what exactly joint commitment and joint intention mean at this stage, we can nonetheless propose that a joint intention is this property that we will argue holds 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. How might we characterize joint intentions?

Proposal 1 Joint intention

> *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 7, something very much like this has been proposed in the literature. Assuming for the moment a tight connection between intention and commitment, the proposed definition does indeed guarantee that the two agents intend to do their respective actions as part of the overall joint undertaking. 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 action. In addition, we can assume that there are no hidden obstacles, in that if both parties did their share, then the joint action would succeed. But even with these strong assumptions, the specification by itself is still too weak, once we allow for a divergence of mental states.

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. In particular, as time passes, one agent may come to doubt that another agent is still committed to the group effort. Doubt negates mutual belief, and hence the joint intention no longer holds. So a second proposal for a definition of joint intention is this:

Proposal 2 Joint intention

> *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 (i.e., is 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, the convoy leader comes to realize privately that she does not know the location of the destination. The intention to lead the others to the destination 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 4.3, 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.

3 Individual Commitment and Intention

Our formal account of individual and joint commitments and intentions [8, 26] 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

[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.

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 (and were).

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.[2] 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 [25], for example), but we will ignore that difficulty here. We assume positive introspection for belief and goal, 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 [21]. 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, conditionally, 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

[2] A fixed point definition is given in [8], and a circular data structure for encoding mutual beliefs is described in [6].

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.

3.1 Individual Commitment

Based on these primitives, we define a notion of individual commitment called persistent goal.[3]

Definition 1 *An agent has a* persistent goal *relative to q to achieve p just in case*

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.

[3] This definition differs slightly from that presented in our earlier work [8], but that difference is immaterial here.

3.2 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 [35, 36]. By doing so, we avoid the notoriously difficult issue of how an intention self-referentially causes an agent to act, as discussed in [35], 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 2 *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 [8], 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.[4] 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.

4 Teams

A team is a set of agents having a shared objective and a shared mental state — without either, there is no unified activity and hence no team. A group of spectators running for cover from a

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

sudden rainstorm may all have a common goal — remaining dry — but there is no coordinated team activity [36]. While the spectators have a common goal they do not have a shared goal. Each has the goal independently of the others, and the success of an individual neither affects nor is affected by the success of any of the other participants. Even where there is coordinated action, without a shared mental state a team does not exist.

4.1 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 3 *An agent has a* weak achievement goal *(or* **WAG** *) 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,[5] 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

[5] 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.

once an agent comes to think the goal is finished, she never changes her mind.[6] 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.[7]

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

Proposal 3 Joint Commitment — *A team of agents have a* joint persistent goal *(or* **JPG** *) 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.*

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.

If an agent discovers the goal has been accomplished or is impossible, or if he discovers the relativizing condition is no longer true, he is allowed to drop the goal. The agent is left with a goal to make his discovery mutually believed by the rest of the team. Consider again the case of a convoy of cars being led to a destination by a driver who thought he knew the way. During the trip, if the driver of a following car realized that they are now at the destination, she should be left with the goal of making sure the lead driver knows that as well. That is she could not just stop without letting the convoy leader know that the task the convoy had been formed to accomplish has been completed. Similarly, if the lead driver comes to realize that he doesn't know the way, he must tell the following drivers before driving off. These clauses also will prevent a convoy member from simply being abandoned by the rest of the convoy if she is forced to stop due to mechanical

[6] For readers familiar with the results in distributed systems theory [20] 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.

[7] 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 $\neg q$ dissolves the commitment regardless of the status of p.

problems. If a car stops without notifying the other drivers, the other drivers will realize that the constraints of the joint action prevent the driver of the disabled car from simply abandoning the convoy, meaning the other convoy members will also stop to determine what has happened.

4.2 Properties of Joint Commitment

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 1 *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 2 *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 [26].

Theorem 3 **Termination of a joint commitment:** *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,[8] 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.

4.3 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.

Proposal 4 **Joint intention** — *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.*[9]

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

It can be shown that joint intention correctly distributes over concurrent actions — if two agents jointly intend concurrent actions A and B, they jointly intend A and jointly intend B (each relative to the overarching joint intention). Each agent individually intends his or her own action, and is committed to the other's action. A more complex relationship holds if A and B are sequential. If the acts are supposed to be executed in "lockstep," in which the agents need to mutually believe each act in the sequence is finished and the next is to start, the joint intention of a sequence will entail appropriate joint intentions for the elements, relative to the overarching joint intention for the sequence. The relativization is essential, just as it is with individual intention with respect to sequential actions, since it means that if the overarching joint intention is given up (e.g., it becomes mutually believed that it is irrelevant), the agents need not still be committed to their parts. Similarly, if one agent believes the other's action is impossible, he can drop the joint commitment even though he is left with the residual commitment to attain mutual belief of impossibility. These and other properties of the definition of joint intention are described in [8, 12]. Finally, one can have joint intentions for partial or underspecified actions (e.g, an action with an unknown part),

[8] 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.

[9] A more precise version of this definition [26] also requires that they mutually know when they started.

using existential quantification over events. Of course, due care needs to be taken with "quantifying into" the various mental states involved in the joint intention definition. The interested reader is referred to [8, 12] for more details.

In summary, based on two primitive mental states, as well as a characterization of actions and time, we arrive at notions of joint commitment and joint intention that can support joint action, and also make fine-grained predictions of the relationships between joint and individual actions and mental states. We now turn to the problem of showing how agents can form and discharge these joint commitments and intentions.

Although the conditions proposed in the analysis rely on mutual belief, strictly speaking, there are ways to acquire mutual beliefs without explicit speech acts (see [5, 28, 31]), such as circumstances in which agents are copresent and can observe one another. However, we will primarily be concerned here with mutual belief acquired via explicit communication.

Given this understanding of joint commitments and intentions, it now remains to be shown how agents can become jointly committed.

5 Communication

We are concerned not only with the case of natural language communication, but also with the case of artificial agents who exchange messages. Although problems of natural language interpretation are important to our research program (e.g, see [10]), they shall be ignored here. Rather, we will assume communication takes place by way of performative utterances [9]. Thus, we assume every communicative act specifies what type of act it is, which is typical for artificial agent communication languages (e.g., KQML [15]). In other work [9], we have demonstrated how the present analysis supports a precise theory of natural language performatives. The theory shows how a present tense declarative natural language utterance constitutes both an assertion of what the speaker is doing, as well as, in the relevant context, a performance of the named illocutionary act. The latter property is true because non-institutional illocutionary acts are defined as attempts, in which the speaker need only have the right beliefs and intentions.

We will assume all agents are willing communicators in that if asked to respond, they do. Moreover, they are sincere in that they never try to get other agents to believe something that they do not want them to know. More formally, we characterize the sincerity assumption as follows:

Definition 4 *Sincerity — An agent is* sincere *with respect to another agent and a proposition p iff when the wants the other to come to believe p, he in fact wants that agent to come to know p.*

We will assume agents are always sincere. With these assumptions, if it is mutually believed that a message was received, it is mutually believed that the named action took place.

Methodologically speaking, we have elsewhere provided definitions of illocutionary acts as action expressions in the dynamic logic described in this paper [10]. Given an event sequence, multiple actions could be said to have been performed (and thus we side with Davidson [14] in analyzing action sentences). But, as argued by Sadock [32], pragmatics is too easy. That is, there are few constraints on what constitutes an illocutionary act definition. We have argued elsewhere for a methodology that involves describing the action in a logical description language thereby enabling

the theorist to derive various properties entailed by those definitions and semantics. In our case, we have used this method to show how our earlier definitions subsume many of the properties of illocutionary acts described by philosophers of language (e.g., [1, 34, 37]). However, because of the fine-grainedness of the logical operators, the logic can often make distinctions that others cannot. As a result, there are sometimes properties of illocutionary acts that are new, and need to accord with intuitions, if not actual dialogues.

5.1 Communicative Action

Generalizing a remark by Searle [34], a communicative action will be treated as an *attempt* by the speaker to convey his mental state. The analysis will assume the usual distinction between illocutionary and perlocutionary effects. More generally, the definition of attempting will distinguish between effects the agent is committed to achieving (which will incorporate the illocutionary effects) and those he wants or hopes to bring about (which will include the perlocutionary effects).

More specifically, let us define an *attempt*.

Definition 5 *An* attempt *using event e by X to achieve p while being committed to at least q is defined to be:*

1. *X believes that both p and q are false*
2. *X wants p to become true as a result of doing e*
3. *X intends (relative to 2) that after X does e, q will be true.*

The upshot of the difference between conditions 2 and 3 is that if q does not obtain, X is likely to try again, since achieving q is a minimal effort requirements. In contrast, condition 3 does not place such a stringent commitment on X with respect to p. For example, in attempting to shoot a basket in basketball, the minimal effort condition is to launch the ball, while the desired result is that it go in the basket.[10]

5.2 Directives

We now consider a request, the prototypical directive action. If the notion of an agreement is to have any import, it should at a minimum hold in the case where one party has requested another to perform an action, and the other party accedes. We claim that a request followed by a commissive action should result in a joint commitment's being formed.

It is important to notice that the classical definitions of requesting (e.g, those in [1, 34, 37]), our earlier definition [11] and those of other computer scientists [27, 29], do *not* have this property. Although there is due consideration for the speaker's *wanting* the hearer to do an act, and perhaps for the speaker's wanting the hearer to form an intention to do so, nothing in those definitions *commits* the speaker to anything. Consequently, there is nothing inconsistent in the speaker's making a request, receiving a confirmation that the addressee will comply, and then deliberately

[10] Philosophers might wish for a tighter causal connection between e, p, and q. If we knew of a precise and semantically well-motivated analysis of causality in a possible-worlds semantics, we would use it here.

making the requested action impossible to achieve. The speaker has merely changed his mind. Since the conditions on requesting were stated using an operator that captures desire, and such desires can be contradictory or, in the case of our **GOAL** operator, simply changed without constraint, no predictions of infelicity can be made.

Furthermore, it is not available to the theorist to use the *intend* operator to say that the requestor intends for the listener to act, since under any reasonable analysis of intention, one can only intend to perform one's own actions. However, a hallmark of teamwork is that all team members are committed to each other's success. One therefore wants to be able to say that the speaker is *committed* to the hearer's action. Our persistent goal operator (**PGOAL**), upon which our notion of intention is built, provides just this capability.[11]

By defining a request to use the persistent goal operator to express the speaker's commitment, we can rule out simple changes of desire. However, this is not yet sufficient, since the speaker may be forced to drop that commitment, for example when he comes to believe that the requested action is impossible. As with the fully formed joint intention, we would still want to say that the speaker needs to inform the addressee of the impossibility of the requested action, even before the requested party has acceded to doing the action.

The substantive claim being made here is that once a speaker has made a request, he is *already* treating the addressee as though he were a team member, even if a reply has yet to be received. As a result, it would be infelicitous for a speaker to change his mind about wanting the requested action done without informing the requested party.

This observation implies that the weak achievement goal (**WAG**) operator should be used in the definition of a request in order to convey that the speaker has a **WAG** that the addressee do the requested action. However, even mutual belief of the speaker's **WAG** and later, of the addressee's **WAG** to do the action, are by themselves insufficient to ensure that a joint commitment has been established. Having a joint commitment implies that a weak achievement goal will persist sufficiently long. However, it is not the case that mutual belief that each party has a **WAG** implies a joint commitment because the **WAG**'s in the joint commitment need to persist. It is sufficient to define a persistent weak achievement goal (**PWAG**) as follows:

Definition 6 *An agent has a* persistent weak achievement goal *(or* **PWAG***) 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* persistent* *goal that the status of p be mutually believed by all the team members.*

[11] In response to our criticism to this effect, Bratman [4] postulated an *intends that** operator that allows an agent to intend that another agent act. Grosz and Kraus [19] base their theoretical apparatus on this concept. However,the semantics of *intends-that*, as well its relationship to commitment, remain murky.

5.2.1 Request

We are now finally in position to define a request.

Definition 7 *A* request *using event e from X to Y to perform action A relative to p is defined to be:*

An attempt using e by X with the

1. *Intention to do e in order to make Y believe that it is mutually believed between Y and X that (before e)*

   X had a **PWAG** *with Y that (after e)*

   - *Y does A, and*
   - *Y forms a* **PWAG** *to do A relative to X's* **PWAG** *that Y do it.*
2. *Goal that, after e*
   - *Y does A, and*
   - *Y forms the* **PWAG** *to do A relative to X's* **PWAG** *that Y do it.*

Based on this definition, we can show the following:

Theorem 4 *Requests imply the requestor is committed —*
If it is mutually believed that a request by X to Y has just happened, and it is mutually believed that X is sincere, and that the event of requesting does not make the requested action true, impossible, or irrelevant, then it is mutually believed that X now has a **PWAG** *that Y do A.*

Proof:
By definition of a request, if it is mutually believed that the request has just happened, then it is mututally believed that X had the intention that Y believe that is mutually believed that X *had* the **PWAG** that Y do A. By the definition of sincerity, it is mutually believed that X only wants Y to believe what is true, so it is mutually believed that X's intention is that X and Y mutually know that X had the **PWAG** tht Y do A. Since X knows (and remembers) what his goals are (were), X could not have that intention if X did not in fact have the **PWAG**.[12] Since it is mutually believed that the request does not perform the requested action, make it impossible, or irrelevant, the **PWAG** in fact persists across the making of the request. Therefore it is mutually believed that X now *has* the **PWAG** that Y do A.

Thus, in sincerely performing a request to do some action, X is committed to making public that he is *already** committed to Y's doing the action and to Y's forming the **PWAG** to do so. Such a commitment to Y's action means that X will screen out incompatible options [3], track the success of the commitment, etc. Thus, merely by requesting Y to do something, X has committed resources, and has taken on the need to communicate should his attitude towards the action change. Y's response could then establish a true joint commitment/intention, or could let X off the hook. To see how, we need at least one more basic communicative action, assertion.

[12] This intention (labeled 1 in the definition) is satisfied in fashion reminiscent of Grice [17] and Schiffer [33] in that it is satisfied by the mutual recognition that it held.

5.3 Assertion

Definition 8 *An* assertion *using event e from speaker Y to X that p is true is defined to be:*

An attempt using event e by Y to X, where

1. *Y intended e to make it the case that Y believe that is mutually believed that before e, Y wanted X to come to believe (after e) Y believes p.*
2. *Y wanted X to come to believe (after e) that Y believes p.*

In asserting p, Y is not trying to get X to believe p, but only to come to believe that Y believes p. As before with request, one can show that if it is mutually believed that Y has asserted p to X, the intention 1 in the attempt definition is again satisfied by means of its public recognition and sincerity. It therefore becomes mutually known that the speaker believes p.

Now, if it is mutually believed that the following sequence of assertions has taken place, then it is mutually believed that some proposition p holds.

Theorem 5 MB establishment —

If it is mutually believed that: an assertion from X to Y that p has taken place followed by an assertion from Y to X that Y believes *p, and X has not privately changed his/her belief about p, then it is mutually believed between X and Y that p.*

Proof Sketch.
After the first assertion, it is mutually believed that X believes p. After the second, it is mutually believed that Y believes p, as long as X doesn't change his/her belief about p before receiving Y's utterance. It therefore becomes mutually believed that p.

One can now define a special case of assertion, which we will call an accede:

Definition 9 *A* Accede *using event e from Y to X regarding action A is defined to be:*

an assertion using event e that Y has formed a **PWAG** *with X to do A relative to X's* **PWAG** *that Y do A.*[13]

6 Establishing and Discharging a Joint Commitment

One can now show the following:

Theorem 6 JPG Establishment —

[13]Of course, a speaker does not in fact need to *say** the content of this assertion (although he could), but the definition incorporates a complex set of mental attitudes. Similarly, the speech act "lie" incorporates a complex pattern of mental states in a single word.

If a request *from X to Y to do A is followed by an* acceding *from Y to X with respect to A, then a joint commitment (***JPG** *) between X and Y that Y do A has been established (relative to X's original* **PWAG** *).*

Proof Sketch. After a request has been performed, it is mutually believed that the speaker X has a persistent weak achievement goal (**PWAG**) that Y do A. After the acceding act, it is mutually believed that Y has a **PWAG** to do A (relative X's). To complete the conditions for establishing a **JPG** , it only remains to show that the agents must mutually believe that A has not been done, and that both parties in fact want it done. These conditions are readily established from the **PWAG**. Note that a subsequent assertion from X that X believes Y has the **PWAG** too is not necessary since agents are introspective about their goals — if Y believes she has a **PWAG** , then she does.

Once a joint commitment has been established, we know from the termination theorem that eventually one of the parties will have the commitment to make the status of the joint goal mutually believed. It is a simple matter to discharge all outstanding commitments undertaken in the **JPG** by making it mutually believed that the joint action has been done, is impossible, or is irrelevant. Thus, considering a joint commitment that A be done, if Y asserts that A has been done, and X asserts that he believes it, then the **JPG** has been discharged. Similarly for asserting that A is impossible or irrelevant.

6.1 Other Speech Acts

Other speech acts should also be able to bring about a joint commitment. For example, if X offers or proposes to Y to do action A for Y, and Y accepts, X and Y should become jointly committed to X's doing A. That means, in particular, that if Y knows of some reason why action A is impossible, he should tell X. Moreover, Y will not take on commitments that would make X's doing A impossible, etc. Thus, the definitions of offer, propose, and accept will need to be revised to use the **PWAG** operator in order to be able to form a joint commitment.

7 Comparison with other approaches

Tuomela [43] has written a definitive work on joint action, relating the concept to that of "we-intentions," norms, and social roles, and culiminating in a "general dynamic theory of society." The present analysis of joint action does not attempt such wide coverage, but rather attempts in the small to link joint action theory with speech act theory.

Rather than be concerned primarily with the necessary and sufficient conditions for joint commitment, we have been focused on what property "binds" the agents together so that they in fact have a joint commitment. Tuomela and colleagues have been concerned with the relationship of "we-intending" to individual intentions, but until recently [43] the conditions under which one can drop joint intentions and commitments have been of secondary concern. The present analysis derives the individual commitment of all team members to the actions of their partners from the semantics of the concepts involved in our definition of joint commitment. Thus, the present analysis also derives as a consequence that team members are required to help others if the success of the

other requires that help. Although such properties may be present in other theories, they are not in fact derived, but rather are stipulated.

A second area of comparison with other theories is that of how joint intentions are formed and discharged. In one major explication, it is claimed that joint intentions require "agreements," either explicit or implicit [43, 73-77]. For example, regarding explicit agreements:

> Agreement-making requires a "communicative" change in the world — a relevant sign indicating agreement. (However, I shall not here attempt to give an exhaustive list of speech acts and other communicative acts which can result in agreement-making and the entailed obligations). [p. 74].

The present paper is squarely addressed at how such speech acts should be defined.

Our earlier analysis was criticized therein in that we only required joint intention to be based on mutual beliefs, rather than on agreements. True enough. That means, we allow for means other than explicit speech acts to arrive at a joint intention. In fact, it could simply be obvious to everyone involved that a joint intention is operative, without there being any communication involved (Gricean or otherwise). Tuomela appears to require something stronger in that he requires there to be at least nonverbal communication:

> Here is one example of a joint action with an implicit agreement: When sitting in our garden we suddenly notice that it is beginning to rain. We quickly form the agreement and joint intention to move our books and papers and other similar items inside. This we may do by shouting to each other something to the effect that it is starting to rain, and you should take care of the chairs while I carry the papers. But we can also bring about more or less the same agreement by giving significant looks. However some kind of signal will be needed in forming a joint intention and acting jointly on that intention: we do not operate merely on the basis of mutual beliefs and without agreement-making. [p. 77]

The operative feature of this account is "significant look." If one were to provide a semantics for a "significant look," it would likely be analyzed as an illocutionary act, resulting in mutual belief because of the reciprocal viewing of the parties. We then agree (in a different sense) with Tuomela on the examples so far, but leave open the possibility that joint intentions can be formed when the appropriate mutual beliefs are in force, without stating how they came to be operative.

Finally, in [12], we pointed out that all analyses of joint intention based on mutual belief fall apart as soon as one of the agents doubts the commitment of the others, or believes privately that the joint action is impossible, irrelevant, or satisfied, etc. This is because the mutual belief operator is false as soon as there is even one embedded negated belief operator. The subsequent analysis in [43] then argued that surely, even if the joint intention is gone,

> "given the agreement-view, we can say that the agreement by A and B that they drive in convoy to B's home... obviously is not yet gone: A is obligated to communicate to B his private belief. If A just speeds away, he surely is not doing his part of the agreement to do X jointly in the sense meant ... [p. 133].

Unfortunately, no precise analysis of agreement (explicit or implicit) has been supplied that derives such an obligation to inform, which is precisely the motivation for our theory. Thus, perhaps the present analysis is elaborating what was meant in [43] in more precise terms.

Grosz and Kraus [19] (hereafter, GK) develop an elaborate axiomatic theory of collaborative plans based on a four-fold distinction among intentions – intending to do a complex action, intending that a proposition be true, and potential intentions to do actions or achieve propositions. In addition, the authors then layer "meta-predicates" on top of these definitions to characterize full and partial individual plans (FIP and PIP, respectively) and full and partial shared plans (FSP and PSP, respectively). An agent's (or a group's) plans are partial when it does not know how to expand an action into subactions terminating in basic actions, or when agents only have a potential intention towards one of the subactions, rather than a full-blooded intention. The axiomatic analysis in GK concentrates on characterizing shared plans both partial and full. The action specification language follows Goldman's [16] and Pollack's [30], and is targeted at hierarchical composition, especially how one action contributes to the performance of another. Within this framework, the authors attempt to capture many of the properties we have derived, including having agents commit to the success of their partners' actions, communicating when joint actions go astray, and providing helpful behavior. Much of the analysis rests on the introduction of the new and controversial *intends-that* operator (following [4]) that is used to characterize one agent's attitude towards another's or a group's actions or plans. Although it is claimed that an agent's intending-that the group performs a collective action *commits* the agent to the success of the group, and implies necessary communication actions, the primitive *commit* operator does not appear in the definition of *intends-that* (though it does in the definition of *intends-to* which captures an agent's intention towards his/her own action). As a result, many of the claims are in need of further support. However, the authors attempt to explain numerous phenomena that we have not tried to capture, especially, how agents further elaborate upon their partial joint and individual plans.

Woolridge and Jennings [45] provide an analysis with a branching-time semantics that generalizes the present one, situating it as one point in a space of possible definitions of joint intention. They address the issue of collaborative problem-solving in four steps: an agent recognizes the need/potential for cooperative action; the agent asks for assistance and forms a team; the team formulates a joint plan ; the team acts to execute joint plan. Thus, this work considers topics that we have not — specifically, we do not treat the first and third of these steps. However, regarding team formation, although the authors state that the formation of the team is done by communicative actions in the style of [10], the specific speech act definitions are not provided.

In [24] Kinny et al. present a formulation of joint intentions in the BDI logic framework. Their definition of join intentions require that each team member to have the intention and that there be mutual belief among the team members that all have the proper intentions. These definitions do not require communication among the team members when a member succeeds or sets aside a joint intention, nor is a requirement for team communication derived from the properties of the defined mental states of the team members. In addition to the definitions, the authors present an algorithm that transforms a plan and a set of team roles into a set of role-plans; communication requirements similar to ours are imposed upon team members by that algorithm.

7.1 Applications.

The theory of joint action has important practical ramifications. As computer scientists, we are interested in applications of those models in a variety of domains. Human-computer interaction may profitably be regarded as human-computer collaboration [38], and theories of collaboration can help to provide new mechanisms for supporting human collaboration. In the field of artificial intelligence, there are numerous attempts to build multiagent systems that can be said to collaborate [23, 22, 40]. For example, Jennings and Mamdani [23] have shown that the present analysis can be used to specify and guide the design of a multiagent system for electric power managment, in which grid-diagnostic agents share a joint intention. As a group, the jointly committed agents waste less time in resolving circuit outages than a collection of "selfishly" motivated agents because they keep each other up to date when the jointly intended actions are satisfied, impossible, or irrelevant.

Tambe et al. [40] are using the joint intention theory discussed here to enhance the performance of teams of aircraft in a military simulation. A typical mission will have a team of helicopters fly to a prearranged jump-off point, carry-out their assigned mission, and return to base. Joint intention theory is used to control individual pilot's actions [41] with respect to the team goal, and to reason about the actions of members of opposing teams [42]. The first implementation of these agents was designed without any explicit model of teamwork; a team of agents was simply a group of agents with common goals. These agents could not form teams that stayed together in the face of changing circumstances or adversity. For example, the teams could not recover if certain team members failed to complete their assigned portion of the mission, some team members could not recognize completion of a portion of the mission if it had not accomplished that subgoal itself, and upon completing a portion of the assigned mission, team members would lose coordination with the rest of the team. When the agents were redesigned to incorporate an explicit model of teamwork, through the addition of operators representing the joint intentions, many of these problems disappeared. The requirement for communication when an agent achieves a goal or comes to believe the goal to be impossible or irrelevant ensured a continued thread of coordination for the duration the mission. Moreover, the commitment to the overall goals of the team provided the flexibility required to continue to perform, to modify or to add goals to an individual's sub-plan when a team member observed a problem in the team's overall plan.

Finally, analyses of how communicative actions can be used to establish and discharge joint intentions can have an impact on the design of interagent communication languages for the Internet [13, 39]. In fact, computer systems are being designed to support simultaneously human-human collaboration and teamwork, human-computer interaction, and interactions among artificial agents [7]. Building a robust and reliable system that incorporates these three components requires a single model of collaboration and teamwork. Without it, the multiplicity of interactions will quickly become too difficult to understand and control.

8 Conclusions

In this paper, we have argued for a definition of joint intention that gives serious consideration to the circumstances in which the team members must drop their joint commitment. Joint intentions

and commitments are built out of the same primitives as individual intentions and commitments. One can easily show that two individuals who have a joint commitment to some goal are each individually committed to that goal. A critical desideratum for a theory of joint action is the ability for agents to form the underlying joint intentions via communication. The paper has redefined a number of basic communicative actions and shown that they can be used to form and to discharge teams. Finally, other work of ours [39] has argued that these definitions of communicative actions can induce the familiar finite-state analyses of dialogue, explaining the nature of the individual states and state transitions. Moreover, it is argued that such communicative action definitions can form the foundation for artificial agent communication languages.

9 Acknowledgments

This research was supported in part by the Office of Naval Research (grant N00014-95-1-1164), and in part by the Information Technology Promotion Agency, Japan, as a part of the Industrial Science and Technology Frontier Program 'New Models for Software Architectures' sponsored by NEDO (New Energy and Industrial technology Development Organization). The results presented here do not reflect the position or policy of either the US Government or the Japanese Government.

References

[1] K. Bach and R. Harnish. *Linguistic Communication and Speech Acts.* M. I. T. Press, Cambridge, Massachusetts, 1979.

[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] M. E. Bratman. Shared cooperative activity. *Philosophical Review*, 101:327–341, 1992.

[5] H. H. Clark and C. Marshall. Definite reference and mutual knowledge. In *Elements of Discourse Understanding.* Academic Press, New York, 1981.

[6] P. R. Cohen. *On Knowing what to Say: Planning Speech Acts.* PhD thesis, University of Toronto, Toronto, Canada, January 1978. Technical Report No. 118, Department of Computer Science.

[7] P. R. Cohen, M. Johnston, D. McGee, S. L. Oviatt, J. Pittman, L. Chen, and J. Clow. Quickset: Multimodal interaction for simulation set-up and control. In *Proceedings of the Fifth Applied Natural Language Processing Conference*, Association for Computational Linguistics, Washington, D. C., April 1997.

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

[9] P. R. Cohen. and H. J. Levesque. Performatives in a rationally based speech act theory. In *Proceedings of the 28th Annual Meeting, Association for Computational Linguistics*, Pittsburgh, Pennsylvania, 1990.

[10] 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.

[11] P. R. Cohen and H. J. Levesque. Confirmations and joint action. In *Proceedings of the 12th International Joint Conference on Artificial Intelligence*, Morgan Kaufmann Publishers, Inc., San Mateo, California, August 1991, pp. 951–957.

[12] P. R. Cohen and H. J. Levesque. Teamwork. *Noûs*, 25(4):487–512, 1991. Also Technical Note 504, Artificial Intelligence Center, SRI International, Menlo Park, California, 1991.

[13] P. R. Cohen and H. J. Levesque. Communicative actions for artificial agents. In *Proceedings of the International Conference on Multiagent Systems*, AAAI Press, Menlo Park, California, 1995.

[14] D. Davidson. Actions, reasons, and causes. In A. R. White, editor, *The Philosophy of Action*. Oxford University Press, 1968.

[15] T. Finin, R. Fritzson, D. McKay, and R. McEntire. KQML as an agent communication language. In *Proceedings of the Third International Conference on Information and Knowledge Management (CIKM'94)*, ACM Press, New York, November 1994.

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

[17] H. P. Grice. Meaning. *Philosophical Review*, 66:377–388, 1957.

[18] 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, pp. 417–444.

[19] B. J. Grosz and S. Kraus. Collaborative plans for complex group action. *Artificial Intelligence*, 86(2), October 1996.

[20] 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.

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

[22] N. R. Jennings. Commitments and conventions: The foundation of coordination in multi-agent systems. *The Knowledge Engineering Review*, 8(3):223–250, 1993.

[23] N. R. Jennings and E. H. Mamdani. Using joint responsibility to coordinate collaborative problem solving in dynamic environments. In *Proceedings of the Tenth National Conference on Artificial Intelligence*, AAAI Press/MIT Press, Menlo Park, California, July 1992, pp. 269–275.

[24] D. Kinny, M. Ljungberg, A. Rao, E. Sonenberg, G. Tidhar, and E. Werner. Planned team activity. In *Artificial Social Systems*, Lecture Notes in Computer Science 830. Springer-Verlag, 1994.

[25] 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.

[26] 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.

[27] D. J. Litman and J. F. Allen. A plan recognition model for subdialogues in conversation. *Cognitive Science*, 11:163–200, 1987.

[28] C. R. Perrault. An application of default logic to speech act theory. In P. R. Cohen, J. Morgan, and M. E. Pollack, editors, *Intentions in Communication*. MIT Press, Cambridge, Massachusetts, 1990.

[29] C. R. Perrault and J. F. Allen. A plan-based analysis of indirect speech acts. *American Journal of Computational Linguistics*, 6(3):167–182, 1980.

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

[31] D. Sadek. Dialogue acts are rational plans. In *Proceedings of the ESCA/ETRW Workshop on "The structure of multimodal dialogue" (VENACO II)*, Maratea, Italy, September 1991.

[32] J. M. Sadock. Comments on Vanderveken and on Cohen and Levesque. In P. R. Cohen, J. Morgan, , and M. E. Pollack, editors, *Intentions in Communication*, MIT Press, Cambridge, Massachusetts, 1990, pp. 257–270.

[33] S. Schiffer. *Meaning*. Oxford University Press, London, 1972.

[34] J. R. Searle. *Speech acts: An essay in the philosophy of language*. Cambridge University Press, Cambridge, 1969.

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

[36] 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.

[37] J. R. Searle and D. Vanderveken. *Foundations of Illocutionary Logic*. Cambridge Univ. Press, New York City, New York, 1985.

[38] C. Sidner and C. Rich. COLLAGEN: When agents and people collaborate. In *Proceedings of the First International Conference on Autonomous Agents*, ACM Press, New York, 1997, pp. 284–291.

[39] I. Smith and P. R. Cohen. Toward a semantics for an agent communications language based on speech-acts. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI'96)*, AAAI Press, Menlo Park, California, 1996, pp. 24–31.

[40] M. Tambe, W.L. Johnson, R. Jones, F. Koss, J.E. Laird, P.S. Rosenbloom, and K. Schwamb. Intelligent agents for interactive simulation environments. *AI Magazine*, 16(1), 1995.

[41] M. Tambe. Teamwork in real-world, dynamic environments. In *Proceedings Second International Conference on Multi-Agent Systems*, AAAI Press,Menlo Park, California, 1996, pp. 361–368.

[42] M. Tambe. Tracking dynamic team activity. In *Proceedings of the National Conference on Artificial Intelligence*. AAAI, AAAI Press, 1996.

[43] R. Tuomela. *The importance of us*. Stanford University Press, Stanford, California, 1995.

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

[45] M. Wooldridge and N. R. Jennings. Towards a theory of cooperative problem solving. In *Distributed Software Agents and Applications (MAAMAW '94)*, Lecture Notes in Computer Science 1069, Springer-Verlag, 1996, pp. 40–53.