SRI International

Technical Note 485 • 1 May 1990

ON ACTING TOGETHER

Prepared by:
Hector J. Levesque
Department of Computer Science
University of Toronto

Philip R. Cohen
Artificial Intelligence Center
Computing and Engineering Sciences Division
and
Center for the Study of Language and Information
Stanford University

José H. T. Nunes
Department of Computer Science
University of Toronto

To appear in Proceedings of AAAI-90

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," and by a contract from ATR International to SRI International. The Toronto authors were supported in part by a grant from the Natural Sciences and Engineering Research Council of Canada.

On Acting Together

Hector J. Levesque*
Dept. of Computer Science
University of Toronto
Toronto, Ont. M5S 1A4
and
Philip R. Cohen
AI Center and CSLI
SRI International
Menlo Park, CA 94025
and
José H. T. Nunes
Dept. of Computer Science
University of Toronto

June 5, 1990

Abstract

Joint action by a team does not consist merely of simultaneous and coordinated individual actions; to act together, a team must be aware of and care about the status of the group effort as a whole. We present a formal definition of what it could mean for a group to jointly commit to a common goal, and explore how these joint commitments relate to the individual commitments of the team members. We then consider the case of joint intention, where the goal in question involves the team performing some action. In both cases, the theory is formulated in a logical language of belief, action, and time previously used to characterize individual commitment and intention. An important consequence of the theory is the types of communication among the team members that it predicts will often be necessary.

1 Introduction

What is involved when a group of people decide to do something *together*? Joint action by a team involves more than just the union of simultaneous individual actions, even when those actions are coordinated. We would not say that there is any *team work* 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 as a single agent with beliefs, goals, and intentions of its own, over and above the individual ones. In this paper, we present a formal model of these mental properties of a group, and

*Fellow of the Canadian Institute for Advanced Research

especially how joint intentions to act affect and are affected by (and ultimately reduce to) the mental states of the participants.

In previous work, we have presented a belief-desire-intention model of the mental states of individuals in which intentions are seen as internal commitments to perform an action while in a certain mental state. 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 about the world or the other agents, as having changeable goals and fallible actions, and as subject to interruption from external events. Whereas the model is sufficient to predict planning, replanning, and communication [2, 1], it does 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. This paper extends our previous work and characterizes joint intentions as shared commitments to perform an action (typically composite) while the group is in a certain shared mental state.

Although we do not explore how these ideas can be applied in computational systems that reason about action, we take the research presented here to be essential groundwork. Some account of joint action is obviously needed to at least describe (or form plans containing) coordinated activities such as jointly lifting a heavy object or writing a paper, as well as pure group activities like games, plays, and dances. The theory also provides a basis for formalizing the type of agreement and commitment underlying legal contracts and treaties. In addition, it allows us to make sense of mundane utterances like "Uh-huh," "OK," "Right," "Go on," "Now" and others, that pepper all our natural dialogues. These can be seen as attempts to establish and maintain the mutual beliefs necessary to the achievement of joint intentions. Indeed, one of the main goals of this theory is to predict and interpret the sort of linguistic activity that arises when agents cooperate. A companion paper [3] will provide details of the application of the theory to dialogue.

In the rest of this paper, we discuss the problem of joint action in general, and then build a formal theory of joint commitment based on the same primitives as the individual case. We then examine some of the properties of joint commitment, and use it to define a simple form of joint intention. Finally, we discuss why it is rational for agents to form joint commitments at all.

2 A Convoy Example

If team behaviour is more than coordinated individual behaviour, how does it work? When is it necessary? When should agents communicate? What should they communicate? These questions are perhaps best answered by considering what would happen in the case of a convoy example *without* the right sort of joint intention.

Consider two agents, A and B, driving somewhere together, with A leading the way until B knows his[1] way home. Assume that both agents have this as their intention, and furthermore, that these individual intentions are mutually believed to hold. In other words, it is mutually known to both agents that each will do his part (if he can) as long as the other agent does likewise. Will this work?

[1] We use masculine adjectives and pronouns throughout, but they should be read as "his," "her" or "its" (for robots), as the reader desires.

This is essentially the model of joint intention proposed by Tuomela and Miller [7] and by Grosz and Sidner [4]. We feel that it runs into difficulties in cases where it is possible for one of the agents to come to believe (privately) that the intention either has been achieved or is impossible to achieve. We assume that in such a situation, a rational agent has no choice but to abandon the goal. For example, if A comes to realize that he was mistaken and in fact does not know where B lives, the intention to lead B must be given up. The trouble is that there is nothing in the agreement to stop A from just speeding away, ignoring a puzzled B behind him. Conversely, if B comes to realize that he now knows his way home, the goal has been satisfied according to him, and there is nothing to stop him from pulling over to enjoy the scenery, without regard to how A might interpret his action.

But the real problem with this characterization of joint intention is that it does not work even if both parties behave with the utmost of consideration for the other, and even if these private beliefs of failure or success do not arise. As long as these beliefs are thought to be *possible*, what can happen is that one agent may come to believe that something like this is happening to the other! For example, if A makes a turn, B could very well (falsely) conclude that A is no longer able to lead him, and so make other plans for getting home. Conversely, if B is forced to pull over (for example, because of difficulties with his car), A could simply conclude that B now knows the way, and continue driving without him. At a deeper level of misunderstanding, even if each agent does not misinterpret the other's actions, he has no way of knowing that the other agent will not misinterpret his! Even if A is indeed prepared to stop in case B does, B might not realize this, and not want to pull over in case A takes it to mean that B now knows his way home. This can continue indefinitely to deeper and deeper levels. Overall then, with the potential for this kind of misunderstanding, even with the best of intentions, nothing is holding the collective behaviour together. Individual intentions do not a convoy make.

So what do we expect from a convoy? Among other things, robustness against misunderstandings like those above: A will signal when it is time to get started, A and B will endeavor to keep each other within sight and not pull over privately, A will not take actions that he believes would render B's intentions impossible to achieve, B will signal when he knows his way, and without such a signal, A will still assume B is following. Of course, A and B do not need to explicitly agree to these actions; they should be consequences of what it means to act together. Before examining a definition of joint action that has these properties, we review the individual case.

3 Individual Commitment and Intention

The account of intention given in [1] is formulated in a modal language that has the usual connectives of a first-order language with equality, as well as operators for the propositional attitudes and for talking about sequences of events: (BEL x p) and (GOAL x p) say that x has p as a belief and goal respectively; (MB x y p) says that x and y mutually believe that p holds; (AGT x_1 ... x_n e) says that x_1 ... x_n are the only agents for the sequence of events e; $e_1 \leq e_2$ says that e_1 is an initial subsequence of e_2; and finally, (HAPPENED a), (HAPPENING a), and (HAPPENS a) say that a sequence of events describable by an action expression a has just happened, is happening now, or will happen next, respectively. An action expression here is built from variables ranging over sequences of events using the constructs of dynamic logic: a;b is action composition; a|b is nondeterministic choice; a||b is concurrent occurrence of a and

b; p? is a test action; and finally, a^* is repetition. The usual programming constructs like IF/THEN actions and WHILE loops can easily be formed from these.[2]

A few comments on how formulas of this language are semantically interpreted. BEL and GOAL are given a possible-world semantics, where a world is modeled as a function mapping each time point to a set of primitive event types (the events happening simultaneously at that point in time). We assume that each agent has perfect introspection about both his beliefs and goals, that beliefs are consistent, and that goals are consistent with each other and with what is believed. Sentences are evaluated at both a world and a current time point on that world. The truth value of sentences with HAPPENED, HAPPENING, and HAPPENS differ only with respect to the position of the current time point: immediately after, straddled by, and immediately before the action, respectively. We will also use the following syntactic abbreviations:

Actions:

$$(\text{DONE } x_1 \ldots x_n\ a) \stackrel{\text{def}}{=} (\text{HAPPENED } a) \wedge (\text{AGT } x_1 \ldots x_n\ a)$$
$$(\text{DOING } x_1 \ldots x_n\ a) \stackrel{\text{def}}{=} (\text{HAPPENING } a) \wedge (\text{AGT } x_1 \ldots x_n\ a)$$
$$(\text{DOES } x_1 \ldots x_n\ a) \stackrel{\text{def}}{=} (\text{HAPPENS } a) \wedge (\text{AGT } x_1 \ldots x_n\ a).$$

Eventually: $\Diamond p \stackrel{\text{def}}{=} \exists e\ (\text{HAPPENS } e;p?)$.

There is something that happens, including the empty sequence of events, after which p holds, *i.e.*, p is true at some point in the future.

Always: $\Box p \stackrel{\text{def}}{=} \neg\Diamond\neg p$.

The wff p is true from now on.

Until: $(\text{UNTIL } p\ q) \stackrel{\text{def}}{=}$

$\forall c\ (\text{HAPPENS } c;\neg q?) \supset \exists a\ (a \leq c) \wedge (\text{HAPPENS } a;p?)$.

Until the wff p is true, the wff q will remain true.

With these definitions in place, we can say what it means for an agent x to be (fanatically) committed to achieving a goal p: he should believe that p is false, but want it to be true at some point, and continue to want it to be true until he believes that it is true, or that it will never be true. Thus we have the following definition:[3]

Definition 1 $(\text{PGOAL } x\ p\ q) \stackrel{\text{def}}{=}$

$$\begin{aligned}&(\text{BEL } x\ \neg p) \wedge (\text{GOAL } x\ \Diamond p) \wedge \\ &(\text{UNTIL } [(\text{BEL } x\ p) \vee (\text{BEL } x\ \Box\neg p) \vee (\text{BEL } x\ \neg q)] \\ &\qquad (\text{GOAL } x\ \Diamond p))\end{aligned}$$

The extra condition q here (which we will occasionally omit) is simply a reason x may have for keeping the goal. It is most often used in an expression such as (PGOAL x p (GOAL x q)) to express a commitment to p as a *subgoal* relative to q.[4] Finally, we define what it means for x to intend to do an action a:

[2] Test actions occur frequently in our analysis, yet are potentially confusing. The expression p?;a should be read as "action a with p holding initially," and analogously for a;p?. Note specifically that an agent can perform these without ever knowing the truth value of p.

[3] This is slightly different from the one appearing in [1].

[4] A better way to do this would be to allow for a dynamically evolving set of *priorities*, and to allow an agent to drop a goal if it is found to conflict with one of higher priority.

Definition 2 (INTEND x a q) $\stackrel{\text{def}}{=}$

(PGOAL x
  (DONE x [UNTIL (DONE x a) (BEL x (DOING x a))]?;a)
  q)

So an agent intends to do an action if he has a persistent goal to have done that action, and moreover, to have done it believing throughout that he was doing it. It is therefore a commitment to do the action *deliberately*. Typically such a goal would arise within a subgoal-supergoal chain as a decision to do an action a to achieve a goal p by getting into a mental state where a would be done knowingly.[5] If the chain is something like

(PGOAL x p) $\wedge$
(PGOAL x (HAPPENED a) (GOAL x $\Diamond$p)) $\wedge$
(PGOAL x (DONE x a) (GOAL x $\Diamond$(HAPPENED a))) $\wedge$
(INTEND x a (GOAL x $\Diamond$(DONE x a))),

then the goal could be given up if the agent discovers that a was performed somehow without his realizing it (or any other goal higher in the chain was achieved).

4 Joint Commitment

How should the definition of persistent goal and intention be generalized to the case where a group is acting like a single agent? Restricting ourselves to two agents here (and throughout), a first attempt at a definition for JPG, *joint persistent goal*, would be to replace belief in the definition of PGOAL by mutual belief, and replace (GOAL x $\Diamond$p) by mutual belief in the goal as in

Definition attempt: (JPG x y p q) $\stackrel{\text{def}}{=}$

(MB x y $\neg$p) $\wedge$ (MG x y p) $\wedge$
(UNTIL [(MB x y p) $\vee$ (MB x y $\Box\neg$p) $\vee$ (MB x y $\neg$q)]
  (MG x y p)),

where

(MG x y p) $\stackrel{\text{def}}{=}$ (MB x y (GOAL x $\Diamond$p) $\wedge$ (GOAL y $\Diamond$p)).

This has the effect of treating x and y together as a single agent, but otherwise leaving the notion of persistent goal unchanged.

However, the definition is not quite right: it will only work in cases where neither agent can come to believe *privately* that the goal p has succeeded or is impossible.[6] To see why, suppose that x alone comes to believe that p is impossible; x must drop the goal, and so (MG x y p) must be false; but this mutual goal was supposed to persist until there was a certain mutual belief, and as there is as yet none, there cannot have been a JPG to start with.

So joint commitment cannot be just a version of individual commitment where a team is taken to be the agent, for the simple reason that the team members may diverge in their beliefs. If an agent comes to think a goal is impossible, then he must give up the goal, and fortunately knows enough to do so, since he believes it is impossible. But when a member of

[5] An alternative but less effective choice would be to blunder about at random, checking periodically to see whether or not the action had been done.

[6] We thank Henry Kautz for this observation.

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 was precisely the problem with the failed convoy discussed above. Rather, we must insist on arriving at mutual belief, that is, at an agreement that the goal is impossible to achieve, before commitments can be discharged. Any team member who discovers privately that a goal is impossible (or has been achieved) should be left with a goal to make this fact known to the team as a whole, which, in effect, is what introspection does in the individual case.

We therefore define the state of a team member x nominally working on p relative to another member y as follows:

Definition 3 $(\mathsf{WG}\ x\ y\ p) \stackrel{\text{def}}{=}$

$$[\neg(\mathsf{BEL}\ x\ p) \wedge (\mathsf{GOAL}\ x\ \Diamond p)] \quad \vee$$
$$[(\mathsf{BEL}\ x\ p) \wedge (\mathsf{GOAL}\ x\ \Diamond(\mathsf{MB}\ x\ y\ p))] \quad \vee$$
$$[(\mathsf{BEL}\ x\ \Box\neg p) \wedge (\mathsf{GOAL}\ x\ \Diamond(\mathsf{MB}\ x\ y\ \Box\neg p))]$$

This form of "weak goal" involves three mutually exclusive cases: either x has $\Diamond p$ as a goal, or thinks that p is true and wants to make that mutually believed,[7] or similarly for p never being true.

If a team is jointly committed to achieving p, the team members cannot assume of each other that they have p as a goal, but only that they have p as a *weak* goal; each member has to allow that any other member may have discovered privately that p is impossible and be in the process of making that known to the team as a whole.

A further possibility (that we will not deal with) is for an agent to discover that it is impossible to make the status of p known to the group as a whole, for example, when communication is impossible. For simplicity, we assume that once an agent comes to think that p is unachievable, he never changes his mind, and that it is always possible to achieve the correct mutual belief. 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.[8]

So the final definition of JPG replaces MG in the last clause of the previous definition by a weaker version:

Definition 4 $(\mathsf{JPG}\ x\ y\ p\ q) \stackrel{\text{def}}{=}$

$$(\mathsf{MB}\ x\ y\ \neg p) \wedge (\mathsf{MG}\ x\ y\ p) \wedge$$
$$(\mathsf{UNTIL}\ \ [(\mathsf{MB}\ x\ y\ p) \vee (\mathsf{MB}\ x\ y\ \Box\neg p) \vee (\mathsf{MB}\ x\ y\ \neg q)]$$
$$(\mathsf{WMG}\ x\ p))$$

where

$$(\mathsf{WMG}\ x\ y\ p) \stackrel{\text{def}}{=} (\mathsf{MB}\ x\ y\ (\mathsf{WG}\ x\ y\ p) \wedge (\mathsf{WG}\ y\ x\ p)).$$

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

[8] 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. In this case, mutual belief in $\neg q$ amounts to an agreement to dissolve the commitment regardless of the status of p.

5 Properties of Joint Commitment

The first thing to observe about this definition of JPG is that like its flawed predecessor, it also generalizes the concept of PGOAL, in that it reduces to the individual case when the two agents are the same:

Theorem 1

$$\models (\mathsf{JPG}\ x\ x\ p) \equiv (\mathsf{PGOAL}\ x\ p)$$

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

It can also be shown that like the previous account, this definition of joint commitment implies individual commitments from the team members:

Theorem 2

$$\models (\mathsf{JPG}\ x\ y\ p) \supset (\mathsf{PGOAL}\ x\ p) \wedge (\mathsf{PGOAL}\ y\ p).$$

To see why x has p as a persistent goal, imagine that at some point in the future x does not believe that p is true or impossible to achieve. Then there is no mutual belief either, and so p must still be a weak goal. But under these circumstances, this means that p must still be a real goal. Consequently, p persists as a goal until x believes it to be satisfied or impossible to achieve.

So if two agents agree to do something, they become individually committed to achieving it. This was stated by Searle as one of the major puzzles of joint intention [6]: given that joint intentions do not reduce to the conjunction of individual ones, where do the individual intentions come from (since ultimately, it is the individuals who act)? In the Grosz and Sidner formulation, joint intentions are *defined* in terms of individual ones. But as we saw earlier, their definition had a drawback given the possibility of private discoveries about the status of the goal. With our definition, however, an agent cannot give up the goal just because he suspects that the other agent has given it up (or suspects that the other suspects that he has, and so on). Until they know the status of the goal itself, they cannot drop it.

So what *does* happen when one agent x discovers privately that p is impossible to achieve?[9] First observe that the agent must now drop the goal of achieving p, and so the PGOAL and the JPG must be false as well. This is as it should be: we would not want to say that the agents continue to be jointly committed to achieving p, since one of them has now given it up.

But (and this is the important point) the fact that there is no longer a joint commitment does not mean that the collective behaviour falls apart. Since there is as yet no mutual belief that p is impossible, we know that the $(\mathsf{WMG}\ x\ y\ p)$ must persist. This means that $(\mathsf{WG}\ x\ y\ p)$ must persist, and $(\mathsf{GOAL}\ x\ \Diamond(\mathsf{MB}\ x\ y\ \Box\neg p))$ must persist too, since $(\mathsf{BEL}\ x\ \Box\neg p)$ is true and will remain true. So although the original JPG no longer holds, because of the UNTIL clause in that JPG, the goal to eventually achieve mutual belief persists until it is achieved. This goal is therefore a PGOAL:

Theorem 3

$$\models (\mathsf{JPG}\ x\ y\ p) \wedge \ldots \quad \supset$$

[9] Similar considerations apply when one of agents discovers that the goal has been achieved, or when the agent thinks that something like this is happening to the other.

$$
\begin{array}{l}
(\mathsf{UNTIL}\ [(\mathsf{MB}\ x\ y\ p) \vee (\mathsf{MB}\ x\ y\ \Box\neg p)] \\
\qquad [(\mathsf{BEL}\ x\ (\Box\neg p) \wedge \neg(\mathsf{MB}\ x\ y\ \Box\neg p)) \supset \\
\qquad\qquad (\mathsf{PGOAL}\ x\ (\mathsf{MB}\ x\ y\ \Box\neg p))])
\end{array}
$$

The ellipsis here is some condition that is sufficient to guarantee that x will not change his mind about the impossibility of p. The simplest and strongest such condition is $\Box[(\mathsf{BEL}\ x\ \Box\neg p) \supset \Box(\mathsf{BEL}\ x\ \Box\neg p)]$, but others are possible.

Similarly, when agent x discovers privately that p has been achieved, the goal of making p mutually believed persists until the JPG is discharged:

Theorem 4

$$
\begin{array}{l}
\models (\mathsf{JPG}\ x\ y\ p) \wedge \ldots \quad \supset \\
\quad (\mathsf{UNTIL}\ [(\mathsf{MB}\ x\ y\ p) \vee (\mathsf{MB}\ x\ y\ \Box\neg p)] \\
\qquad [(\mathsf{BEL}\ x\ p \wedge \neg(\mathsf{MB}\ x\ y\ p)) \supset \\
\qquad\qquad (\mathsf{PGOAL}\ x\ (\mathsf{MB}\ x\ y\ p))])
\end{array}
$$

To summarize: once a JPG has been established, if one of the agents comes to believe that the goal has been achieved or is impossible, the individual commitment to achieve p is replaced by a new commitment to make the status of p mutually believed.

This has two very important consequences. First, this PGOAL to attain mutual belief predicts that *communication* will take place as this is typically how mutual belief is attained, unless there is co-presence to begin with. To satisfy a contract, in other words, it is not enough to satisfy the agreed upon goal (or to find it to be unsatisfiable), one must be prepared to *show* the other that it has been satisfied. This explains why contracts normally have concrete *deliverables*, and why it would be strange to have a contract requiring one of the parties to merely think about something.

Second, 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 communication of the status of the goal. We do not merely require the agent to work on the goal while he believes the other agent to be doing the same, since in many natural cases, one agent will lose track of what the other is up to. Instead, he must work on a goal until there is mutual belief regarding the status of the goal. As we discuss below, what makes this at least reasonable is the fact that an agent can rely on the other to let him know if he is wasting his time on an impossible goal.

Let us reexamine the convoy example in the light of these theorems. First, we introduce some (simplistic) notation:

$$(\mathsf{know\text{-}way}\ y) \stackrel{\text{def}}{=} \exists z\ (\mathsf{BEL}\ y\ (\mathsf{way\text{-}home}\ y\ z)).$$

$$(\mathsf{done\text{-}convoy}\ x\ y) \stackrel{\text{def}}{=} (\mathsf{DONE}\ x\ y\ (\mathsf{WHILE}\ \neg(\mathsf{know\text{-}way}\ y)\ [(\mathsf{leads}\ x);(\mathsf{follows}\ y)]))$$

So (know-way y) is intended to say that y knows his way home, and (done-convoy x y) says that x and y have just done the iterative action consisting of x leading and y following (whatever that means) until y knows his way home.

If x and y are jointly committed to doing the convoy action, we can show that if y comes to know his way home, he cannot simply ignore x and go home; he remains committed to making it mutually believed that he knows his way home (for example, by signalling):

Theorem 5

$$\models (\mathsf{JPG}\ x\ y\ (\mathsf{done\text{-}convoy}\ x\ y)) \wedge \ldots \quad \supset$$

```
(UNTIL [(MB x y (done-convoy x y))
          ∨ (MB x y □¬(done-convoy x y))]
       [(BEL y (know-way y) ∧
              ¬(MB x y (know-way y))) ⊃
          (PGOAL y (MB x y (know-way y)))])
```

All that is needed to show that this theorem follows from Theorem 4 is the fact that the worlds where y knows his way home are precisely the worlds where the WHILE loop has just ended (after perhaps zero iterations). More complex properties of the convoy depend on a joint commitment to more than just the proper completion of the WHILE loop, as in joint intention, which we now turn to.

6 Joint Actions

Given the notion of joint commitment, we define JI, *joint intention*, as the obvious generalization of individual intention:

Definition 5 (JI x y a q) $\stackrel{\text{def}}{=}$

```
(JPG x y
     (DONE x y
            [UNTIL (DONE x y a)
                   (MB x y (DOING x y a))]?;a)
     q)
```

So joint intention is a joint commitment to do an action while mutually believing (throughout the execution of the action, that is) that the agents are doing it. Space permits us only to sketch broadly some of the implications of this definition.

Typically, the a in question will be a composite action involving parts to be done by each agent alone. For example, a could be $a_x;a_y$ or $a_x||a_y$, where a_x is some action to be performed by x alone, and similarly for a_y. Since both parties are committed to getting *all* of a done, both parties care about the other's actions and so will not intentionally do something that would make them impossible. If one agent does his part, but sees that the other agent has difficulty doing his, this definition predicts that the first agent will want to *redo* his part, to get the whole thing right. If there are turns to be played, each agent will have to make sure that the other knows when it is his turn, perhaps by a signal of some sort. In fact, neither agent is committed to his part of the bargain in isolation; individual intentions to do one's part do *not* follow from a joint intention to do a composite action. Acting alone could very well be ineffectual (as in lifting a piano) or worse (as in the "coordinated attack" problem [5]). An agent that discovers that his partner's action is impossible may refuse to do his part, even if it remains possible to do so.

The other feature of joint intention is that the action needs to be performed in a certain shared mental state. The main consequence of this is that it predicts that communication will be used to ensure that both parties are aware throughout the execution of a that the action is being done. In addition to signals that transfer control when taking turns (noted above), one would expect to see a signal at the start of the action (like "Ready"), and various reassuring confirmation signals (like "Uh-huh") to make sure the initial mutual belief does not dissipate over time [3].

7 Conclusion

In our previous work, we discussed individual actions and intentions in terms of the rational balance agents maintain among their beliefs, goals, and commitments. We conclude here by discussing why we feel it is *rational* for agents to enter into joint commitments. Our account predicts (perhaps counterintuitively) that an agent will persist in trying to achieve a goal even if he happens to believe the other agent is in the process of informing him of why he had to give it up. Why is this persistence a better strategy than letting the other agent do all the work or dropping the goal as soon as there is uncertainty in the air? There are two reasons. First, the agent knows that he will eventually be told by his partner if he is working on a futile goal. If in fact he is doing more than he strictly needs to, he at least knows that his partner is committed to rescuing him. Second, if he were to take the more conservative strategy and quit immediately, or even if there were suspicions to that effect, the collective behaviour would fall apart and doom the project. This is not unlike the classical Prisoner's Dilemma problem where if both agents fail to cooperate and choose the *locally optimal* strategy, the global result is unacceptable to both parties. It is the mutual commitment to a non-conservative form of behaviour that binds the team together.

Of course the *real* problem in interpersonal affairs is trying to arrive a truly shared commitment. If one of the parties is suspicious of the goals of the other, then by our definition, there is no joint commitment, even if the other party thinks there is. International treaties are most often predicated on *verifiability*, that is, on ways to assuage suspicions as they arise. But suspicion at any level (even a belief that the other party believes that *you* are suspicious) implies that there is no mutual belief that the goals of the treaty are shared. Thus, there is no joint commitment, and like the failed convoy example, the treaty will not be robust in difficult situations. Sad, but true. What it takes to build *trust* in potentially adversarial situations is perhaps the single most delicate aspect of multi-agent interaction. Our account of joint commitment obviously does not provide criteria for avoiding deception or for recognizing true commitment when it exists; but it does state precisely what one is trying to recognize, and what believing in a commitment amounts to.

8 Acknowledgments

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," and by a contract from ATR International to SRI International. The Toronto authors were supported in part by a grant from the Natural Sciences and Engineering Research Council of Canada.

References

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

[2] 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*. M.I.T. Press, Cambridge, Massachusetts, in press.

[3] P. R. Cohen, H. J. Levesque, J. Nunes, and S. L. Oviatt. Task-oriented dialogue as a consequence of joint activity. in preparation, February 1990.

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

[5] J. Y. Halpern and R. Fagin. A formal model of knowledge, action, and communication in distributed systems: Preliminary report. In *Proceedings of the 4th ACM Conference on Principles of Distributed Computing*, New York City, New York, 1985. Association for Computing Machinery.

[6] 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, in press.

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