Chapter 21

Artificial Intelligence and Collective Intentionality: Comments on Searle and on Grosz and Sidner

Jerry R. Hobbs

1 *Comments on Searle*

I'm a hard-core, unreconstructed AI researcher, and so I naturally read Searle's paper from a hard-core AI point of view. That point of view forces one to recast his argument into other terms, but once the recasting is done, one's reaction is, "Of course, who would ever have thought otherwise?"[1] This point of view, however, leads one to see collective intentionality in a larger framework and also leads one, for independent reasons, to disagree with his final conclusion.

First, let me outline the hard-core AI point of view, as I see it. An *agent* (a person or a successful AI program of the future) is a planning mechanism. The agent has *beliefs* and *goals* and uses its beliefs, especially beliefs about what kinds of things cause or enable what other kinds of things, to construct *plans* for achieving those goals. To eliminate spurious mysteries like the mystery of where goals come from, we can think of an agent as going through the world continually modifying, in light of the new beliefs it acquires through perception and reasoning, a plan to achieve the single goal "I thrive."[2]

A goal is simply a proposition, or rather a logical expression representing a proposition, that the agent tries to make true. The proposition can be anything, not just actions on the part of the agent; there are no restrictions on its predicates or arguments. Just as pick-up(I, BLOCK3, NOW) is a possible goal, so are push({JOHN, I}, CAR5, NOW) and win(S.F.GIANTS, World-Series, 1996). The goals can represent events in which I myself am the agent and am able to perform directly, events I can bring about in concert with others, and events I have little control over. The goals can be immediate, like the first two, or they can be long-term, like write(I, Great-American-Novel, Before 2002).[3]

The agent's beliefs are likewise logical expressions that are manipulated in complex ways by the planning mechanism and other processing of the agent's. The agent's beliefs accord with perception, and he or she or it acts as though the beliefs were true. Of particular interest are the agent's beliefs about what causes or enables what else, since they are the material out of which plans are built.

A plan consists of a top-level goal, such as "I thrive," and decompositions of this goal into subgoals and these subgoals into further subgoals, and so on. (For convenience, I'll usually refer to the subgoals simply as goals.) The events described in the subgoals of a goal must be believed in aggregate to cause or enable the event described in the goal. High-level, long-term goals can and usually do have lower-level, more immediate goals, so write(I, Great-American-Novel, Before 2002) might ultimately spawn the subgoal buy(I, PENCIL, NOW). A plan is thus a representation of believed causal relations among events. This is not all, however. To be a plan, it must terminate in actions the agent is capable of executing directly at the appropriate time, such as move(I, ARM1, NOW), or in conditions that will be true at the appropriate time anyway, such as coast-down(CAR5, HILL2, NOW).

An agent moves through the world continually elaborating, modifying, and executing a plan for achieving the goal "I thrive." The agent at any given time may execute an action, discover from the environment's response that it was unsuccessful, replan, perhaps even abandoning some high-level subgoals of "I thrive," execute another action, be satisfied of its success, take some time to develop its long-term plan, and so on.[4]

Many of the notions of folk psychology can be recast into the terms of this framework. Routine or habitual behavior can be viewed as the execution of precompiled plans. Intentions can be viewed as goals, providing we don't take our intuitions about the use of the English word "intend" too seriously. Future-directed intentions (see Bratman, chapter 2 of this volume) are high-level goals expressing conditions in the future. Intentions-in-action are low-level goals to be executed now. Desires can be viewed as beliefs about the efficacy of one event or condition causing another, where both are tied ultimately to the goal of "I thrive." Admittedly, desires don't *feel* like other beliefs, but they can be modeled formally as such, and doing so has the advantage of placing them in a framework that allows them to be analyzed, as most if not all of our desires can be. For example, I have a desire around two o'clock every afternoon to eat a candy bar. But this is not undecomposable; it is directly related to giving me pleasure, which is related in some complex way with my thriving, but it also gives me a burst of energy that makes me more productive in the afternoon and is thus related to other, high-level goals. Reactions to events in the environment, even intense and immediate reactions like jumping into a river to save one's child, can be viewed formally as examples of rapid replanning in light of new information from the environment, even though, again, they don't *feel* like replanning.[5]

That's the whole story.

Well, almost. There is one more piece to the puzzle. Among people's and other agent's beliefs about causal forces in the world are their folk

theories of human
hiking up the side
boulders hurtling
Camino and see a
away, we don't v
intention. We beli
sidewalk. We have
in various circumst
intend to do. Thes
many events with
two roles of inte
"belief," and "plan
agents is described
intend is one part
set of logical expr

Now, from the
have collective int
as goals in its plan
collectives engage
an agent, me for ex
the car ("We push
"You push car," w
"I push car" is a g
could be further d
"You push car" mi
in Searle's example
may be that you d
discuss below. In
action, I have to
us that we have t
belief that the even
time.

An agent's colle
and the correspond
mutual belief is to
have the right inte
times.

Now let's look
have a plan of pur
goal "I help huma
all MBAs that if x
interests as a way

theories of human action based on the notion of "intention." If we are hiking up the side of a mountain and we suddenly see a bunch of one-ton boulders hurtling down, we panic. If we are walking on the sidewalk of El Camino and see a bunch of one-ton automobiles hurtling past, not ten feet away, we don't worry. The difference between these two situations is intention. We believe the drivers of the automobiles intend to keep off the sidewalk. We have theories about what kinds of action people will intend in various circumstances, and we believe that people usually do what they intend to do. These theories, or belief systems, allow us to predict a great many events with significant reliability. It is important to keep distinct the two roles of intention I have introduced. The terminology of "goal," "belief," and "plan" is the metalanguage with which the cognitive life of agents is described. The folk theory of human action involving the predicate intend is one particular set of beliefs among others, that is, one particular set of logical expressions, that the agent has.

Now, from the hard-core AI point of view, Searle's statement that we have collective intentions translates into a statement that an agent can have as goals in its plan, logical formulas whose predicates describe actions that collectives engage in and whose agent argument is such a collective. Thus, an agent, me for example, can form a plan with the goal that you and I push the car ("We push car") that decomposes into the subgoals "I push car" and "You push car," with the proper temporal relations between these actions. "I push car" is a goal that can be executed directly (let's say, although it could be further decomposed into individual bodily movements). The goal "You push car" might be achieved by a variety of means. It may be that, as in Searle's example, it's already true. It may be that I have to ask you, or it may be that you do it simply out of your "sense of community," a notion I discuss below. In addition, for me to believe I'm engaging in collective action, I have to have the belief that it is mutually believed by both of us that we have the same plan. This mutual belief is what gives me the belief that the events you are responsible for will happen at the appropriate time.

An agent's collective intention is then a plan with such a collective goal, and the corresponding belief in the mutual belief in the plan. The role of the mutual belief is to assure each agent in the collective that the other agents have the right intentions and will thus perform the right actions at the right times.

Now let's look at the MBA examples. In the noncollaborative case I have a plan of pursuing my own selfish interests as a way of achieving the goal "I help humanity." Moreover, I believe it is mutually believed among all MBAs that if x is an MBA, then x will have plan of pursuing x's own interests as a way for x to help humanity. Let $P_1(x)$ be the plan

x helps humanity.
|
x pursues self-interest.

Then my plan $P_1(I)$ is

I help humanity.
|
I pursue self-interest.

In addition, I have the belief

$$\mathit{mb}(\mathit{MBAs},(\forall x \in \mathit{MBAs})\,\mathit{has\text{-}plan}(x, P_1(x))).$$

In the collaborative case I have a plan P_2 whose goal is "All of us MBAs help humanity," one of whose subgoals is "I pursue my own selfish interests" and whose other subgoals are "x pursues x's own selfish interests" for every other MBA x:

We help humanity.
I pursue self-interest. A pursues self-interest. B pursues self-interest. ...

Moreover, I have the belief that it is mutually believed among all of us MBAs that every MBA has this plan:

$$\mathit{mb}(\mathit{MBAs},(\forall x \in \mathit{MBAs})\,\mathit{has\text{-}plan}(x, P_2)).$$

Thus, collective intention, when properly recast into a language of beliefs, goals, and plans, understood in their technical senses, is straightforward to characterize in the standard AI planning formalism. Moreover, we can now see collective intentionality as one part of a much larger picture. Nearly everything we do is intricately intertwined with events beyond our direct control happening in the environment. We depend on the universe behaving in certain ways. When I run my pen across the page, I depend on the physical properties of the pen to leave a trace of ink behind it. Our actions are often no more than mere nudges that enable these processes. I might push my car on a slight decline to overcome static friction, so that gravity can accelerate it to the desired speed. When I flip a light switch to turn on a light, I am only performing the last small required action in an enormous chain of events and conditions that society and physics have prepared for me.

Collective actions are a special case of this mesh of the agent with the world at large. The agent has as a goal that the team execute a pass play, for example. A subgoal of this is that he block the defensive end; he can take care of this on his own. Another subgoal is that the quarterback throw the ball at the proper time and in the proper way; belief in mutual belief in a common plan assures our agent that this will happen at the appropriate time. Another event in the plan, yes, as a subgoal if the plan is carried out to that level of detail, is that the ball follow the right sort of parabolic arc; the agent's beliefs about physics, naive or otherwise, assure him of that.

All of this has been so easy it makes one suspect there are really some deeper issues involved here. And indeed there are. The events and conditions represented in the agent's beliefs, goals, and hence plans, have to be framed out of concepts made available by the agent's implicit or explicit folk theories, or systems of belief, about various classes of phenomena. To analyze collective action, therefore, we need better explications and formalizations of people's folk theories of social entities and social action, including some concepts which Searle in this and other writings has made significant contributions to our understanding of, such as commitment, which creates mutual belief in a collective plan among agents, and responsibility, which holds each agent to his or her part.

This point, in fact, is relevant to an argument Searle makes against "talk about group minds, the collective unconscious, and so on." It seems to me the criterion for the acceptability of this vocabulary is the same as it is for any other vocabulary: What work does it do for us? Although I personally don't expect things to turn out that way, the vocabulary would be perfectly acceptable if description in its terms yields a successful theory of social action, say, as successful as our theories of personal action couched in terms like "belief," "goal," "plan," and perhaps "intention." It would be especially acceptable if we were able to articulate its terms with the terms of a successful theory of personal action.

Searle closes section 1 of his paper with an argument that collective intentionality introduces a new and perhaps disturbing fact—"that I can not only be mistaken about how the world is but am even mistaken about what I am in fact doing." But this is not a feature of collective intentionality per se. Rather, it is a feature of the larger framework of hierarchical planning in which I have tried to embed collective intentionality. If while driving I take my foot off the accelerator and I think I'm coasting down a slight hill in my car, whereas in fact I'm on a slight upgrade being carried along by inertia, I'm mistaken about what I am doing. In both this case and Searle's, I have a high-level goal involving in its implementation agencies other than myself. I believe correctly that I am doing my part to achieve this goal. I believe mistakenly that the other agencies are doing their parts.

By embedding collective intentionality in the larger framework of our interactions with causal forces in the world in general, I don't mean to imply that there are no distinctions to be made between people and other natural entities, like rivers and stones. Of course people are very complex, and we interact with other people in very complex ways that we don't experience with rivers and stones. But rivers are also very complex, and we interact with them in complex ways that we don't experience with people and stones.

Section 2 of Searle's paper is a valiant struggle toward an adequate hierarchical planning formalism, without benefit of the relevant literature. In fact, his treatment of hierarchical planning, as well as his emphasis on the importance of notation and his adherence to materialism and methodological solipsism, are so much in tune with the prevailing views and concerns in AI that one wonders why in other papers he has attempted to cast himself as a critic of AI.

In any case, there is one issue he is concerned with in section 2 that is not captured, at least explicitly, in the hard-core AI planning framework. He wants to preserve in his notation for collective intention the self-referential character of intention that his notation for individual intention makes explicit. When one intends to perform an action, one moreover intends that this very intention will cause that action. My favorite illustration of this aspect of intention is from Theodore Dreiser's *An American Tragedy*. The hero, Clyde Griffiths, stands up in a rowboat with the oar, intending to knock his pregnant girl friend into the water and drown her. He slips, the oar hits her and knocks her into the water, and she drowns. He intended to knock her into the water, he knocked her into the water, but the intention didn't cause the knocking.[6]

This self-referentiality has certainly not been made explicit in any planning formalism. But is it implicit? I think one can argue that it is. In an AI framework the self-referentiality means that an agent's goal that an event *e* occur itself plays a causal role in the occurrence of *e*. But this is the case. Because the agent is a planning mechanism, once a goal is formed it persists, unless the plan is modified. The goal is decomposed into executable actions, and when the appropriate time for these actions arrives, they are executed. This is simply the way the planning mechanism works. Thus, the goal does play a causal role in its own achievement. Clyde Griffiths modified his plan at the last instant, so his prior goal to kill his girl friend was not the direct cause of the oar hitting her.

Moreover, agents are aware of this self-referentiality. When an agent plans, he knows the plan has to bottom out in events that will happen at the appropriate time. One of the principal ways events can happen at the appropriate time is by being actions the agent is capable of executing directly. One does not normally plan to have lucky accidents at the appropriate times; it is too undependable; too many such plans fail.[7] The agent knows, in the form of some folk theory or other, that he is a planning mechanism, so he knows that if he decomposes his goals to executable actions and as long as he does not modify his plan, his goals will play a causal role in their own achievement. Of course, the agent must be able to know when the goal to perform a directly executable action has caused that

action. This h
the agent can
they do the m
has become (
feedback and
sometimes is
causing the a(
either that his
being moved
moving, it wa
as part of his j

In section
community" v
core AI point
for an agent t
stick closely t
set of other
members of t
cooperative a
ance that this
associated wi
science gener
would bet th
community is

It might be
of mutual bel
have a sense
don't want to
we people ha
beliefs and th
bottom of the
and, unbeknov
situation that
everyone else
community. A
use their know
when a pickpo
More comple
behaves in a
behavior; wait
assuming othe
there is fullblo
of the sort exe

action. This happens because of feedback from effectors. Via this feedback the agent can distinguish between when the effectors are moved and when they do the moving, even though the trajectory is the same, and the agent has become or has always been aware of the correlations between this feedback and the process of planning. When the feedback is absent, as it sometimes is in injury or under medication, our sense of the intention causing the action falters.[8] In Clyde Griffiths' case, he would have known either that his arms that were carrying the oar toward his girl friend were being moved rather than doing the moving, or that if they were doing the moving, it was as part of a suddenly new plan to brace himself rather than as part of his just abandoned or postponed plan to kill her.

In section 3 Searle raises the very interesting issue of the "sense of community" we can feel and its relation to collective action. From a hard-core AI point of view, a first attempt at a characterization of what it means for an agent to have a sense of community would be this (and here I try to stick closely to what Searle has said). I have a sense of community with a set of other agents if I believe it is mutually believed by me and the members of that set that we are all "actual or potential members of a cooperative activity," and moreover I have a long-term goal of maintenance that this remain true. This may not explain the warm feeling often associated with some of my senses of community;[9] AI and cognitive science generally have little to say about warm feelings. Nevertheless, I would bet that this or something like it explains the ways a sense of community is manifested in action.

It might be objected that to characterize a sense of community in terms of mutual belief has things the wrong way around. Social animals must have a sense of community and they surely don't have mutual beliefs. I don't want to get deeply into biology. But there is a range of experiences we people have that shows the correspondence between complexity of beliefs and the depth of the sense of community they engender. At the bottom of the scale is the situation in which I'm rushing toward a shelter and, unbeknownst to me, so are many other people. One level above is the situation that transpires after it has gradually dawned on all of us that everyone else has the same goal. This already constitutes a weak sense of community. At a higher level is the situation in which some of the people use their knowledge of the common goal to determine their own actions, as when a pickpocket plans where he will place himself in the coming crush. More complex social behavior is possible when everybody in a crowd behaves in a dependable way in order that others may depend on their behavior; waiting one's turn to get through the entrance to the shelter, assuming others will be fair, could be explained in this manner. Finally, there is fullblown collective action, with its mutual belief in a common plan, of the sort exemplified by a previously choreographed convergence on the shelter. At the lower end of this scale, people can act on the basis of how they believe others will act. But human behavior is complex and these beliefs are often wrong. At the higher end of the scale, mutual belief, commitment, and a sense of responsibility make the behavior of others more reliable, allowing each agent to risk and hence achieve more individually, and allowing the collective as a whole to accomplish more. When we participate in a dance, we put ourselves at risk. When a ballerina executes a grand jeté, she needs to know her partner will be there to catch her.[10]

A sense of community is thus not part of a mysterious or unanalyzable "Background." Rather, it consists of certain beliefs and goals, certainly in the technical senses of the terms, and, I think, even in the ordinary senses of the terms, although my ordinary English may have become hopelessly tainted by my work in AI. Searle says that "these are not in the normal case 'beliefs'," likening them to "my stance toward the objects around me and the ground underneath me ... being solid, without my needing or having a special belief that they are solid." Such an appeal to intuition won't work on a hard-core AI researcher like me. My intuition is that it is obvious that I have a belief that the ground is solid, again not only in the technical sense of "belief" but in the ordinary sense as well. When someone out of a background sense of community spontaneously offers to help me push my car, I think there *is* a rule that can be stated explicitly that he or she is acting in accordance with. In fact, it has a name. It's called the Golden Rule.[11]

How do we acquire a sense of community? The same ways we acquire other beliefs and goals. We learn from experience and by being told that particular groups have the potential for cooperative activity, and we adopt in the ongoing plan with which we approach daily life the goal of maintaining this situation, because we have determined that it helps us thrive. People are certainly predisposed to acquire senses of community with those with whom they are or might be engaged in cooperative activity, just as people are predisposed to acquire beliefs about up and down and the solidity of physical objects. There might even be an abstract sense of community hard-wired in, just waiting to be instantiated with various convenient specific communities. But none of this changes the basic outlines of the AI account. It is no more significant than the hardware-software distinction in computer science. It is still just beliefs and goals.

Searle argues against the position, presumably that of Habermas, "that collective behavior presupposes communication, that speech acts in conversation are the 'foundation' of social behavior and hence of society." He wants to argue that there must be a prior sense of community for either collective behavior or communication to occur. This is a chicken-and-egg problem, and the solution is same as in the chicken-and-egg problem. A sense of community, collective behavior, and communication evolve together, both in the evolution of the species and in the development of the individual. We are born knowing how to suckle and desiring to cuddle. Engaging in this collective behavior establishes a small-scale sense of community, which makes possible further collective behavior and communication. Language is learned, enabling us to learn the rules and conventions of quite complex social activities. And so on, until we arrive at the complex creatures we are.

2 *Comments on Grosz and Sidner*

In all work in AI there is a powerful vision that informs the research and can be used as a vocabulary for analysis, and there are the rather simpler and weaker procedures that are actually implemented. The latter are scaled-down versions of the former that one resorts to just because of the difficulties of implementation that every AI programmer is aware of. Grosz and Sidner have chosen to begin their paper with a review of the implemented versions of previous work, rather than of the grander visions. From this perspective, they are quite right in saying the previous work has not dealt with collaborative planning. Previous researchers have dealt with complex collaborative and even more complex conflictual planning on an informal level (for instance, Bruce and Newman 1978; Hobbs and Evans 1980; Wilensky 1978), but previous *implementations* have made the simplifying assumption that only a single agent is doing the planning and acting. Though Grosz and Sidner do not offer an implementation, they do propose a formalism, and if it were successful, it would represent an advance in the study of planning. Unfortunately, it is not successful, and Searle's paper and my critique of Searle's paper are relevant to the question of why it fails. The crucial issue turns out to be the issue of whether a goal is an action or an event.

In single-agent planning this question does not matter very much. Suppose the goal is for the agent A to move BLOCK1. We can express this by a wff representing an event:

move(A, BLOCK1).

We can express it by a wff representing a state that is the end state of the event:

moved(A, BLOCK1).

Or we can express it by a lambda-expression representing the action whose performance by the agent would constitute the event,

λx [move(x, BLOCK1)],

or, as it is more commonly written,

move(BLOCK1).

In single-agent planning translation from any one of these forms to any of the others is trivial. Unfortunately, in multi-agent planning this is no longer the case. Since actions can be performed by more than one agent, we need to be explicit about who is doing what. It is no longer possible for goals to be actions (or at least actions that leave the agent unspecified).

In much of the previous work Grosz and Sidner cite, the agent was made explicit in goals and in operators for effecting goals. In her very fine work, Pollack, dealing only with single agents, was able, as a simplification, to use actions as goals. Grosz and Sidner have tried to carry this simplification over to multi-agent planning, with unfortunate consequences. This apparently was not mere negligence on their part, but a conscious decision. They say, "The desire to provide an appropriate account of imperative utterances (that is, one that did not depend on the notion of agent's intending for another agent to intend to do some action) was a primary motivation for SharedPlans" (section 5.1). I take it that the key intuition here is that one can intend only one's *own* actions. This illustrates one of the pitfalls of taking the ordinary English word "intention" too seriously, especially in the context of a philsophical tradition that has emphasized intentions-in-action over future-directed intentions.

As a consequence of this decision, Grosz and Sidner's formalism is less than perspicuous. We find, for example, expressions of the following sort:

GEN-Simultaneous[lift(foot-end) & lift(keyboard-end), lift(piano), S1&S2].

S1 and S2 are agents. The first problem is whether we are to take the symbol & to mean conjunction, as it usually does, and if so, what the conjunction of S1 and S2 means. Is it, for example, the same as the conjunction of S2 and S1, or is there some significance to the order of the "conjuncts" in the first and third arguments of GEN-Simultaneous? Does the order mean that necessarily S1 is lifting the foot-end and S2 the keyboard-end? It is just not clear who is doing what. If we do a string match of this expression with its definition, we may surmise that the authors use ampersands where they should have commas. Otherwise, the expression has no compositional semantics and is very misleading.

Similarly, and crucially, in the expression

INT(S2, BY(lift(keyboard-end), lift(piano))),

who is lifting the piano? Surely not S2. And S1 is mentioned only in the next conjunct. And what does BY mean? Lifting the keyboard-end does not by itself effect the lifting of the piano. Does BY mean something like "contribute to"? We are just not told.

This may seem a mere notational quibble, but notational difficulties often signal underlying ontological difficulties, and that is the case here.

The use of a notation like lift(foot-end), rather than lift(S1, foot-end), must have come from a desire to preserve actions on the part of a single agent as the only possible object of an intention, and as Searle shows in his paper, this simply won't do. In fact, under the most reasonable interpretation of the unclear portions of Grosz and Sidner's formalism, the definition they give of a SharedPlan is an axiomatization of Tuomela and Miller's attempt to reduce collective intention to individual intention-in-action. As such, it falls prey to Searle's MBA counterexample.

Let us demonstrate this in detail. Suppose S1 and S2 are two of the MBAs who intend to help humanity by serving their own interests. We will show that they have a SharedPlan to help humanity by showing that they satisfy Grosz and Sidner's definition of a SharedPlan at the beginning of section 5. Satisfying Clause 1, they mutually believe that S1 is going to serve his own interests and that S2 is going to serve her own interests. Satisfying the most reasonable reading of Clause 2, they mutually believe that simultaneously serving their own interests will generate helping humanity. Satisfying Clause 3, they mutually believe they each intend to serve their own interests. Satisfying Clause 4, they mutually believe that S2 intends to help humanity by serving her own interests, and similarly for S1. Satisfying Clause 5, they each intend to serve their own interests. Satisfying Clause 6, S2 intends that by serving her own interests, she will—accomplish? contribute to?—the helping of humanity, and similarly for S1. They therefore satisfy Grosz and Sidner's definition of SharedPlan, but as Searle has pointed out, they are not engaging in collective behavior.

Precisely what is missing in this account is what Searle has tried to captured in his notion of we-intentions, and what I have argued is simply a matter of having as a goal an action whose agent is a collective. There is no intention that *we* do something. Moreover, this is a possibility that Grosz and Sidner explicitly reject in their belief that agents can only intend their own actions.

The differences among the four accounts can be summarized succinctly as follows. Tuomela and Miller attempt, by defining collective intention, to reduce it to individual intention-in-action and mutual belief. Grosz and Sidner posit a special operator for collective intention, which they call SharedPlan, but by defining it they also attempt to reduce it to individual intention-in-action. Searle simply stipulates the existence of we-intentions and says they must be, not *defined*, but *implemented* in terms of individual intentions-in-action. I argue that Searle's we-intentions are simply "intentions that we . . ." and as such are particular examples of the manifold ways in which agents plan their intricately meshed interactions with the world in general.

The difficulties with Grosz and Sidner's action notation have another unfortunate effect. They are forced to invent what in this paper is only a few, but must eventually be a myriad, new GEN relations—one for when the actions are done simultaneously, one for when the times overlap a bit, one for when they are sequential, and so on. It is easy to see where this leads. If the collective action is something involving many agents and precise timing, such as Searle's example of a pass play, we would end up with predicates like GEN-49ers-pass-play-1987-36A. A more reasonable notation would have agents and times as explicit arguments. For example, suppose aggr is an operator that takes two wffs representing events and returns the aggregate event, and AGGR maps two individuals into the aggregate of the two. Then one could, say, with much more clarity,

$$\begin{aligned}&\text{GEN(aggr(lift(S1, keyboard-end, t1), lift(S2, foot-end, t1)),}\\&\qquad\text{lift(AGGR(S1, S2), piano, t1)).}\end{aligned}$$

Here it is clear who is doing what and when. No GEN-Simultaneous operator would be needed, and no further GEN operators would be needed for other temporal relations among the actions. The relations would be captured by relations among the time arguments of the predications representing the events. The more explicit and perspicuous notation would force them to abandon their intuition that agents cannot have as goals actions on the part of other agents, but so much the better, for one would then see shared plans as one example of how agents plan their actions to mesh with the larger fabric of what will happen in the rest of the world.

Instead of this elaboration on the interaction of collaborative action and time, it would have been interesting to see an examination of collaborative planning where the actions are discursive in nature, and not primarily physical events in the world such as lifting pianos. I have in mind the sort of behavior that occurs often in a conversation when the participants are so much in synch that one feels they are carrying out a single shared discourse plan. This is especially notable in Falk's (1980) examples of dueting and in Wilkes-Gibbs's (1986) examples of completions. Similar behavior is exhibited in Clark and Wilkes-Gibbs's negotiations of referential expressions (see chapter 23 of this volume).

Two more points are worth commenting on. First, Grosz and Sidner were eager to eliminate the "master-slave" assumption underlying their previous work. This is the your-wish-is-my-command rule that says the hearer adopts the speaker's goals. But they have not replaced it with anything else that can do a similar job. There is rule CDR1 that says that generally a statement of a desire is a request for help, and there is rule CDR2 that says that generally agents try to achieve the shared goals in ways their coparticipants desire.[12] But there is no rule that allows us to go from G1's desire for a shared plan to G2's adoption of that plan. In their analyses of the simple dialogues, they are silent about how that move is made. From Utterance 1 of Dialogue 1 we can infer that S1 wants them both to lift the piano, and from Utterance 2 of Dialogue 1 we can infer that S2 has accepted the shared plan. But *why* S2 accepted it is left a mystery. This is certainly safe, and perhaps even appropriate if we are only eavesdropping on the conversation and want to know what shared plan of action has finally been arrived at. But if we are modeling S2's behavior, then we would definitely have to say something about why S2 chose to adopt the shared plan. Even if we are modeling S1's behavior, we need some way of explaining why S1 thought expressing his desires would lead to S2's adopting them for her own. Of course, the real story here is *much* more complicated than a "master-slave" assumption. We go through all sorts of maneuvers to get others to adopt our desires as their own. It would have been interesting to see an explication of how some of this happens. In its absence, this account no longer explains why a speaker would bother to express his or her desires at all.

Incidentally, Grosz and Sidner attribute the prevalence of the master-slave assumption to the fact that most work on planning has involved only single utterances. This is possibly true in part, for in the case of single-utterance interpretation such a stance is justified. We can decompose the problem of responding to an utterance into the problem of the hearer's determining what to do if he or she were to do everything the speaker desired, and the separate problem of deciding whether or not to actually do it. In investigating the first of these problems, one might as well adopt the master-slave assumption. A more likely source for the prevalence of the assumption, however, is in the fact that much of the early work on multi-utterance dialogue focused on expert-apprentice dialogues, where the apprentice was assumed to be willing to do everything the expert desired.[13]

My second point is illustrated in Grosz and Sidner's account of Dialogue 1. They go through a curious bit of analysis. "From the context in which Utterances 3 and 5 are uttered, the participants can infer that the mentioned actions are seen to participate in a generation relationship with the desired action. That these actions together are sufficient is implicit in Utterances 7 and 8. S1 and S2 can now infer that the generation relation exhibited in Clause 2 holds." But this is surely not right. The real, and obvious, story is that they both already know what it takes to lift a piano and they know that it is common knowledge in our culture.[14] The general problem is that Grosz and Sidner, in this work as elsewhere, do not give sufficient emphasis to the importance of background knowledge in the interpretation of discourse.

Notes

1. This is actually a rational reconstruction. The historical facts are this. I read Searle's description of the problem, and I thought, the answer is such and such. Then I turned the page and read that some people thought the answer was such and such, but here was a counterexample. I thought, "Oh." *Then* I was driven to analyze the problem from

a hard-core AI point of view, which led me to say, "Who would ever have thought otherwise?"

2. Which may or may not have "I survive" as a subgoal.
3. This notation, by the way, does not constitute a serious proposal for an adequate knowledge representation language. Serious proposals do exist.
4. Some in AI may object to my calling this "*the* hard-core AI view." I'm sure, however, they would object more strongly if I claimed originality. I believe it is the point of view at least implicit in Sacerdoti 1977. It is the view expressed in Hobbs and Evans 1980. Rosenschein (1981) formalized Sacerdoti's work along rather different lines. He dismissed hierarchical planning in a few sentences and built his model around the temporal sequence of the agent's actions. This seems to me a mistake. By contrast, I have dismissed the temporal sequence of actions in a few sentences and built my account around the hierarchical character of the planning. It is my view that this hierarchicality is fundamental to the ability of finite creatures like us to operate in an information-rich world, whereas time is just one of many things we reason about and negotiate our way through.
5. I have just introduced a technical vocabulary, and there's a real chance for misunderstanding here. Philosophers often view their task as one of probing ordinary English words, like "belief," "goal," "plan," and "intention," to their conceptual roots, taking their intuitions about these words very seriously. By contrast, an AI researcher plunders ordinary English for lexical items to turn into technical terms. In the framework I have presented, "belief," "goal," and "plan" are technical terms, and "intention" is not. These technical terms are defined precisely and henceforth are unrelated to the ordinary English words except etymologically. In particular, beliefs, goals, and plans need not be conscious, and we can have as goals events such as relaxing and chatting that wouldn't be thought of as especially goal-directed in ordinary folk psychology. Explanations are constructed using these technical terms, and the explanation either succeeds or fails to account for the relevant evidence. Actually, this is not quite true. The ordinary English words are used for their suggestive power. One thinks of an intuitively satisfying explanation in ordinary English terminology, based on folk psychology. One then tries to translate this account into technical terminology. But whether the technical version succeeds as an explanation has nothing to do with its intuitive roots. It depends only on whether it accounts for the behavior in question and meshes well with a theory that accounts for much more of the behavior.
6. He was convicted of first-degree murder nevertheless.
7. Many people buy lottery tickets, but not so many people buy a Mercedes on credit *because* they've bought a lottery ticket.
8. I need not point out that today, in the age of the computer, unlike in the age of Descartes, the causal influence of the mental on the physical is no more mysterious than the conversion of electromagnetic energy into mechanical energy.
9. For example—and this is true—I am sometimes moved to tears reading the *California Driver's Handbook*.
10. It would be interesting to know what various social animals are willing to risk where they must depend on the behavior of others. My sense is that dogs, for example, rarely enter into situations they can't bail out of in fairly short order if their support fails. It's not easy to train a dog to ride a seesaw.
11. A typical sort of argument Searle and others have made elsewhere against the possibility of explicating this background as a set of formal rules operated on by complex inferential processes goes like this: Suppose you went into a restaurant and ordered a hamburger, and the waitress brought you a purple felt pillow six feet in diameter in the shape of a hamburger. How could a formal system be able to deduce this was inappro-

priate? I
rule like
shape o
in argun
manner
useful e
shapes,
a reason
about th
of resear
interpret
zation. I
discover
to the de

12. This is
negotiat
13. A better
what kn
apprentic
interview
usually c
14. Actually,

References

Bruce, Bertram
233.
Falk, Jane (1
Berkeley
Hobbs, Jerry
Cognitive
Rosenschein,
Seventh I
Sacerdoti, Earl
Wilensky, Rol
ment of
Wilkes-Gibbs,
Doctoral

priate? It could never have anticipated such a situation, and so it could not have had a rule like, "Waitresses don't bring customers purple felt pillows six feet in diameter in the shape of a hamburger when they order hamburgers." All of the examples I have heard in arguments like this have been derivable, at least informally, in a fairly straightforward manner from ordinary facts that it is quite reasonable to suppose people have and find useful every day—facts about what kinds of things are edible, the normal colors, shapes, sizes, and material constitutions of common objects, how much food constitutes a reasonable meal, and so on. There are of course problems in how to state and reason about this knowledge, problems that are being addressed vigorously by a large number of researchers. Another argument rests on the necessarily context-dependent nature of interpretation, but this is an argument against only the most naive attempts at formalization. It has been the major thrust of AI work in natural-language processing to discover precisely how context, represented as a structured knowledge base, contributes to the determination of interpretations.

12. This is still pretty slavish behavior. It is more common in equal partnerships to negotiate the means as much as the ends are negotiated.
13. A better way to look at the expert-apprentice dialogues is in terms of who possesses what knowledge. The expert is in possession of the general principles, while the apprentice possesses knowledge of the specific situation. This is true in doctor-patient interviews and in service encounters in general. It is curious that general knowledge usually confers power.
14. Actually, anybody *I* could lift a piano with could probably lift it alone.

References

Bruce, Bertram C., and Dennis Newman 1978. Interacting plans. *Cognitive Science* 2, 195–233.

Falk, Jane (1980). The conversational duet. In *Proceedings of the Sixth Annual Meeting*, Berkeley Linguistics Society, Berkeley, CA.

Hobbs, Jerry R., and David Andreoff Evans (1980). Conversation as planned behavior. *Cognitive Science* 4, 349–377.

Rosenschein, Stanley J. (1981). Plan synthesis: A logical perspective. In *Proceedings of the Seventh International Joint Conference on Artificial Intelligence*, Vancouver, B.C.

Sacerdoti, Earl (1977). *A structure for plans and behavior*. New York: American Elsevier.

Wilensky, Robert (1978). Understanding goal-based stories. Research Report 140, Department of Computer Science, Yale University, New Haven, CT.

Wilkes-Gibbs, Deanna (1986). Collaborative processes of language use in conversation. Doctoral dissertation, Department of Psychology, Stanford University, Stanford, CA.