A Cognitive Model for Collaborative Agents

George Ferguson and **James Allen**
Department of Computer Science, University of Rochester, Rochester, NY, USA
{ferguson,james}@cs.rochester.edu

Abstract

We describe a cognitive model of a collaborative agent that can serve as the basis for automated systems that must collaborate with other agents, including humans, to solve problems. This model builds on standard approaches to cognitive architecture and intelligent agency, as well as formal models of speech acts, joint intention, and intention recognition. The model is nonetheless intended for practical use in the development of collaborative systems.

Introduction

The goal of this paper is to define a cognitive model of collaborative planning and behavior that provides a foundation for building collaborating, communicating agents. We are particularly interested in situations that involve human and non-human (robotic or software) agents working together.

There are several challenges involved in developing such a model. We need to show how agents, who are only able to act individually, can nonetheless plan and perform activities jointly. We also must show how the agents communicate to coordinate all their joint planning and acting, as well as learning to perform new tasks. There is significant prior work formalizing joint activities and shared plans (*e.g.*, (Cohen and Levesque 1990; Grosz and Kraus 1996; Clark and Schaefer 1987; Clark 1996; Subramanian, Kumar, and Cohen 2006)), collaborative problem solving and mixed initiative planning (*e.g.*, (Ferguson and Allen 2006)) and models of communication based on speech act planning (*e.g.*, (Allen and Perrault 1980; Cohen and Perrault 1979)). However these models focus more on formal aspects of belief states and reasoning rather than how agents behave. Other work, such as COLLAGEN (Rich and Sidner 1997) and RavenClaw (Bohus and Rudnicky 2009) focus on task execution but lack explicit models of planning or communication. The PLOW system (Allen et al. 2007) defines an agent that can learn and execute new tasks, but the PLOW agent is defined in procedural terms making it difficult to generalize to other forms of problem solving behavior. Our goal is a model that builds on the theories, accounts for collaborative behavior including planning and communication, and in which tasks are represented declaratively to support introspection and the learning of new behaviors.

Architecture of a Basic Agent

We start by drawing ideas from cognitive architectures, especially ICARUS (Langley, Choi, and Rogers 2009) with its commitment to hierarchical task structures, and from "reactive" (or "BDI") agents such as PRS (Georgeff and Lansky 1987; Ingrand, Georgeff, and Rao 1992) and the "Basic Agent" (Vere and Bickmore 1990). These models assume that the agent knows a set of hierarchical tasks (whether predefined or learned) and that acting involves perceiving the world, identifying which tasks to perform and then decomposing tasks as necessary until the agent identifies the next action it will perform. The agent thus is in a never-ending cycle of perception, goal selection, planning and execution. This can be summarized in the following simple algorithm for the basic agent:

Loop Forever:

1. Process perceptions
2. Identify new problems and tasks
3. Decide which task to focus on next
4. Decide what to do about this task
   Do we know what to do next?
   (a) Yes: do it
   (b) No: Figure out what to do (one-step expansion/decomposition using operators/templates). If successful, expand the task, otherwise abandon subtask and mark as failed.

In practice, we use a more complex multi-nested loop where the agent reconsiders its goals less often and spends more time planning and executing. But these details are not important for this paper.

Note that while this is a model of a single cognitive agent, it is possible to generate communicative speech acts using the planning models developed by Perrault, Allen, and Cohen (1978; 1979; 1980). Such an agent can communicate but cannot create collaborative plans. We'll see this in the first examples we present.

Domain

For concreteness, we'll use a simple domain where agents can push boxes around rooms of a building. Although obviously an abstraction and very simple, this domain emphasizes many key features of real-world problems:

- It naturally scales to more complex problems: with more rooms, doors, and more complex connectivity; by varying colors, shapes, sizes, *etc.*, and adding more complex constraints on goals.
- It can include goals that require joint action (*e.g.*, heavy boxes requiring two agents to push)
- Durative and simultaneous actions are built-in properties of the domain.
- External events can be naturally included, for example by introducing "external" agents that move boxes.
- Observability is a controllable parameter of the domain: *e.g.*, Agents might not be able to identify heavy boxes from their appearance, but have to experiment.

This "Box World" serves as crisp, extensible abstraction of the types of domains and problems that require multi-agent collaboration. It is also very amenable to simulation, in support of visualization and experimentation in scenarios involving human and robotic or software agents working together. Although not the focus of this paper, we are implementing a feature-rich, interactive, immersive simulated environment based on videogame engine technology (specifically "first person shooter," or "FPS" games). The first-person perspective of FPS games allows humans to participate in the world along with automated systems, providing a true testbed for human-centered multi-agent collaboration.

Intentional State and Individual Action

To explore the model further, we need to further elaborate on the agent's intentional state. The agent has a set of tasks it has committed to performing—these are its intentions. As a task is executed, the agent acquires beliefs about the progress of the task and about the world, including whether subtasks have succeeded or failed. To keep the development simple in this paper, we will generally focus on the intentions, and depict them graphically as in Figure 1(a), which shows agent A1 having a task of delivering two boxes, BOX1 and BOX2, to a different room, ROOM2. As indicated by the ? for the agent role of this task, it is not specified which agents should deliver the boxes. Nor does this intention specify how exactly the task should be performed.

We will assume there are two agents, A1 and A2. If A1 were to decide that it could deliver both boxes, it might plan to do so and execute that plan. This would be a straightforward application of the basic agent model, in which the agent successively chooses tasks, decomposes them until it identifies an executable action, and then performs that action.

Instead, for the rest of these examples, we suppose that A1 has decided that it should deliver BOX1 and that A2 should deliver BOX2 (Figure 1(b)). The arrows indicate the decomposition or "in order to" relationship, but we elide many of the representational details. Note that this decomposition of the main task is part of A1's *private* intentional state, which now asserts that A1 intends to accomplish task T *by means of accomplishing* subtasks T1 and T2. Thus the structure of the task is part of the intentional state, just as was done in (Allen and Perrault 1980).

A1, following our model of agent behavior, can make progress on this task. Since it is the agent of T1, it can execute the task (perhaps after refining it further first). However when it comes to T2, agent A1 is stuck. If it cannot communicate with A2, all it can do is wait and hope that A2 will perform T2. Of course since A2 is not privy to A1's private plan, it might do something different (or do nothing). In some cases A1 might be able to reason that A2 already plans to do T2 for some reason, but this is unlikely in general, particularly if A2 is a human agent.[1]

If A1 and A2 can communicate, however, then the situation is more interesting. A1 can adopt the intention of getting A2 to do T2 (Figure 1(c)). Using the standard model of planning speech acts (Perrault, Allen, and Cohen 1978; Cohen and Perrault 1979), this can be accomplished by A1 requesting that A2 perform T2, and A2 accepting that request. These sub-tasks are labeled with "CA" to suggest that they are communicative acts (a generalization of speech acts). Now, A1 can further execute this (private) task by performing the REQUEST, which it does with an utterance like "*Would you please move box 2 into room 2?*"

Suppose A2 is willing to do this, and so responds "*Ok.*" Since this paper is not about natural language understanding, we ignore all language understanding and speech act interpretation issues, but see our prior papers (*e.g.*, (Hinkelman and Allen 1989)). Figure 1(d) shows the situation from A1's perspective after initial interpretation of the utterance as A2's acceptance of A1's request.

Key to our model is that part of the agent's perceptual processing includes the introduction of new tasks to react to other agent's communication. In this case, A1 acquires the task T6 of reacting to the utterance. Performing this task triggers a built-in interpretation process that determines how the utterance fits the agent's beliefs and intentions, and hence how the agent should react (as in (Allen and Perrault 1980)). This *intention recognition* process will be illustrated more fully in subsequent examples. In this case, it is a simple matter of matching the interpretation of A2's utterance CA6 with A1's expectation that A2 will accept (CA5).

With this interpretation, task CA5 is marked as completed, which in turn leads to T3 being completed (that is, A1 has got A2 to intend to do T2). Thus, A1 now believes that A2 will (eventually) do T2. Thus, A1 concludes that the overall task, T, will be completed successfully, even if it simply waits. Indeed, acquiring such beliefs is the reason for performing speech acts under the standard model.

However it is worth noting that although this way of accomplishing the main task involves both agents and com-

[1] Much current work on cooperation in multi-agent systems is based on modeling agent behavior using probabilistic models such as DCOPs (Mailler and Lesser 2004; Modi et al. 2005) and POMDPs (Roth, Simmons, and Veloso 2005; Wu, Zilberstein, and Chen 2009). These are generally planned in advance, rather than collaboratively and incrementally during problem solving. They also do not seem well-suited to modeling interaction involving humans. Some recent research advocates a return to the BDI framework (Taylor et al. 2010).

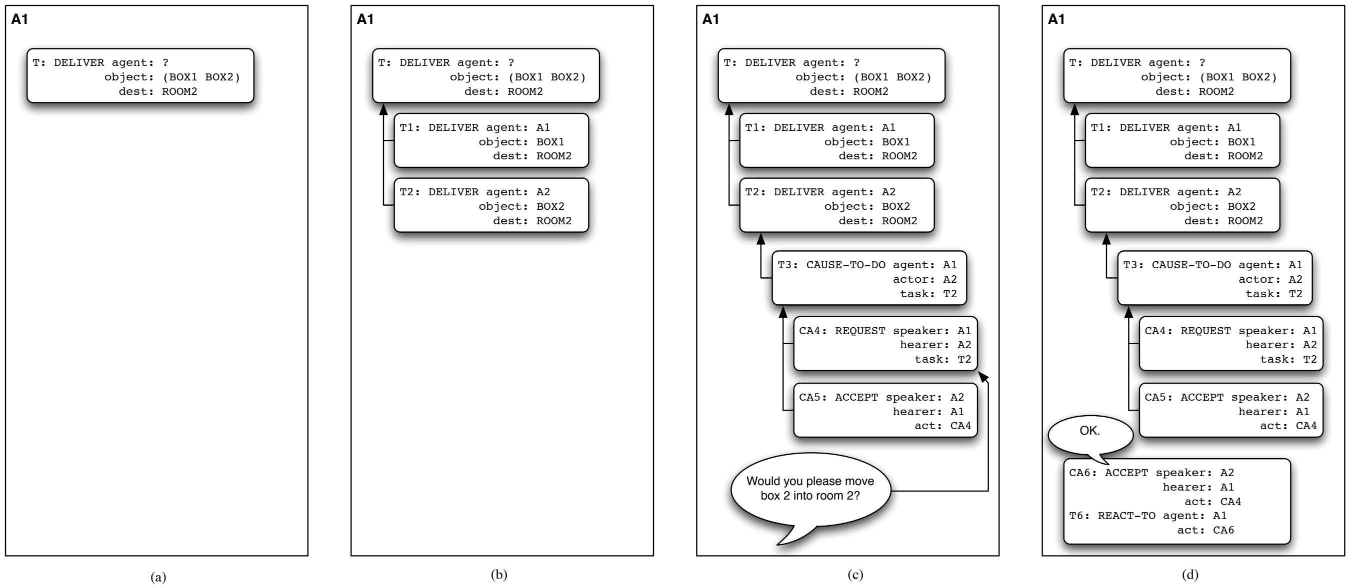

Figure 1: Non-collaborative behavior

munication between them, the planning is nonetheless *non-collaborative* since the entire development has been in terms of A1's private tasks. Thus for example, A1 cannot know whether A2 is doing T2 in order to do T, or for some other reason. They do not have a *shared* task (corresponding to a joint intention as in (Cohen and Levesque 1990; Levesque, Cohen, and Nunes 1990)). This leads us to extend the basic agent model to include the mechanisms necessary for collaboration.

The Basic Collaborative Agent

There are two general approaches we could take to extend the basic model. We could, for instance, modify the basic agent algorithm to account for agent in collaboration. But this seems unmotivated. Even when collaborating, an agent can only perform individual actions, and all its reasoning is still necessarily private reasoning. We do not believe our cognitive architecture changes fundamentally just because we need to collaborate.

But something significant does change, and that is in the agent's intentional state. Besides having private beliefs and tasks, the agent must have a notion of shared tasks (involving shared beliefs and joint intentions). The key point is that beliefs about shared tasks only come about as the result of agreements with the other agent.

Let's return to the example again, where the agent starts with the goal in Figure 1(a). Rather than doing private planning as before, the agent might decide to enlist the other agent's help in defining the plan, *i.e.*, create a shared plan. Note that we treat this as is just another way of decomposing the original goal: the agent plans to achieve task T by making T a shared task, and it does this by getting the other agent to agree to do task T jointly.

This is shown in Figure 2(a). First, the figure illustrates that there are now two task spaces from A1's perspective: its private tasks and the tasks it shares with A2. Second, A1 cannot unilaterally decide that a task is shared. Instead, A1 adopts the intention of accomplishing T by agreeing with A2 to make T a shared task. This action is the first example of what we call a *collaborative problem solving act* ("CPSA" in the figures). A small inventory of such actions represent the agent's knowledge of how to collaborate. An initial analysis of such CPS actions can be found in (Allen, Blaylock, and Ferguson 2002).

Now, the task of agreeing on something with another agent can be decomposed in several ways depending on the acting agent's beliefs and their beliefs about the other agent. For some additional description of this process with somewhat different terminology, see (Ferguson and Allen 2006). In the example in Figure 2(b), A1 decides to accomplish the agreement (T7) by proposing the shared task to A2 and expecting them to accept the proposal. A1's execution of this task leads it to perform the proposal T8, resulting in something like "*Let's work together to move the boxes*" (ignoring the details of natural language generation).

Suppose A2 responds "*Ok.*" As before, initial interpretation of this utterance identifies it as accepting the proposal, and A1 acquires the task of reacting to the utterance (T10). The interpretation triggered by T10 is as before, matching the interpretation of the utterance (CA10) with A1's expectation CA9. This results in CA9 being marked as successfully completed, and thus T7 being completed. The effect of AGREE-ON tasks like T7 is the acquisition of a belief about the shared task, in this case it is the new belief that A1 and A2 share the goal of moving the boxes (Ts).

As noted previously, a crucial element of our model is that *only collaborative problem solving actions can change the contents of shared task spaces.* In this case, A1 recognizing the successful agreement to adopt a new shared task results in the task being added to A1's shared task space as Ts (Figure 2(d)). That is, A1 intends to perform Ts in order to perform T (its private task). This situation illustrates an

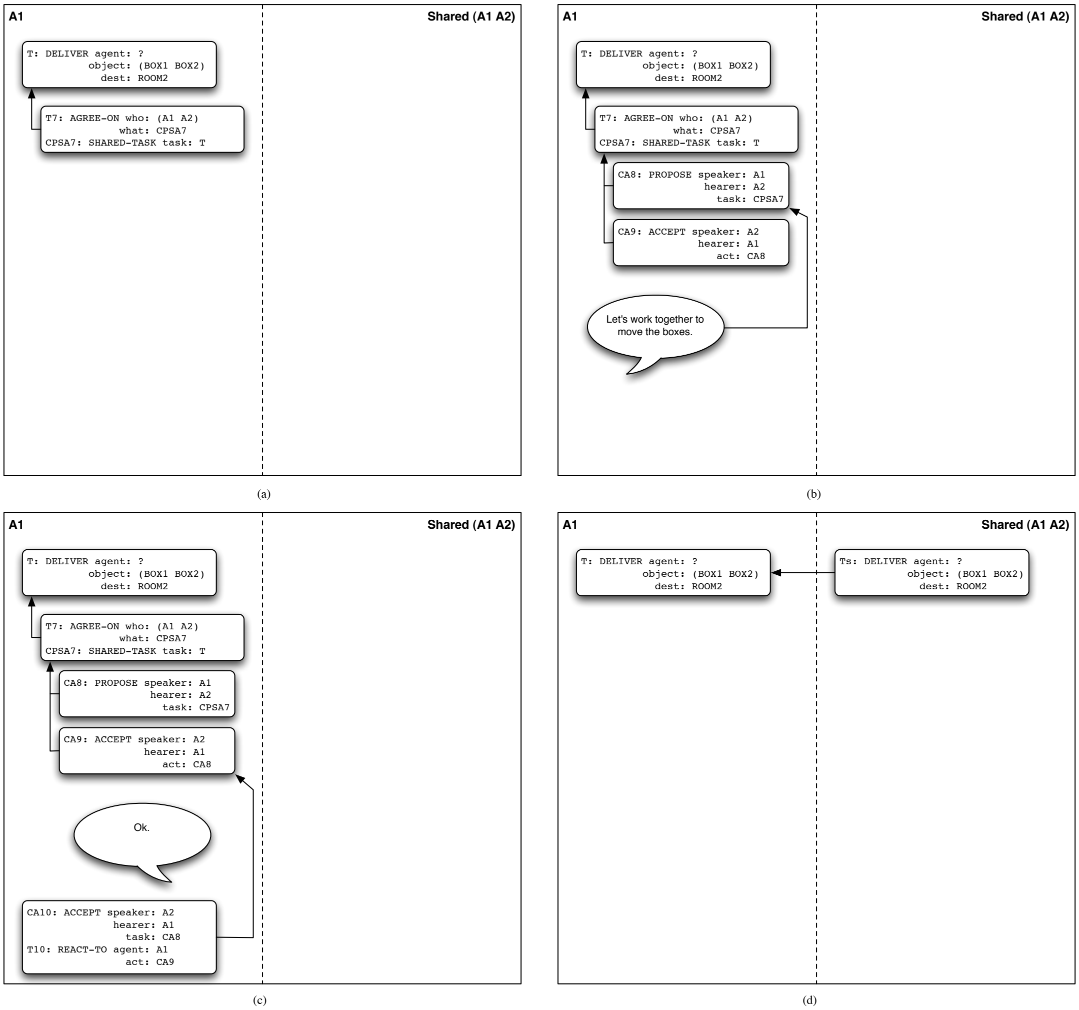

Figure 2: Adopting a shared task

important point about our representation. The task spaces reflect A1's different beliefs and intentional state (private and shared), but the task structures cut across such spaces. It is not the case that we have separate tasks in the shared space and private space. We have a one task T (rooted in private space) that A1 intends to accomplish by means of task Ts in the shared space. Also, as noted above, A1 could not have unilaterally adopted Ts as a shared task. Instead the *collaborative* process of agreeing to do something together leads to a shared task (or joint intention).

Collaborative Planning

So now A1 has established a shared goal with A2. It could simply decide to decompose that shared task privately. That is, it could plan and perform private sub-tasks to accomplish Ts, just as in the previous two cases (Figures 1 and 2). This behavior is not ruled out by our model, but it would probably be a bad strategy. Most of the time when working with someone else, it is more productive to agree on how to proceed. Joint commitment to choices puts stronger constraints on behavior and leads to better teamwork (Cohen and Levesque 1991; Tambe 1997).

Nonetheless, all A1 can do is private planning, even to achieve shared goals. It can, however, choose to accomplish Ts by agreeing with A2 that the way to do it is for A1 to deliver BOX1 (T1) and A2 to deliver BOX2 (T2). This is shown in Figure 3(a) (definitions of T1 and T2 are in Figure 1(a)). To do this, we need to reify the problem solving acts that are used in planning (as in (Allen et al. 2007)). Here, we use the problem solving act DECOMP to stand for the decomposition operation in HTNs. We have an inventory of problem-solving acts that capture all possible operations an HTN-style planning might do: introduce tasks, decompose tasks, bind parameters, abandon tasks/subtasks, modify tasks, *etc*. An agent could use these actions to meta-plan about its own private tasks as well, and such an ability to introspect is critical for learning new problem-solving behaviors. But the point here is that the same problem-solving acts can be proposed as part of an AGREE-ON task in order to perform collaborative planning. Using the same technique from the previous example, A1 proposes a decomposition of the task to A2: "*I'll deliver box 1 and you deliver box 2.*"

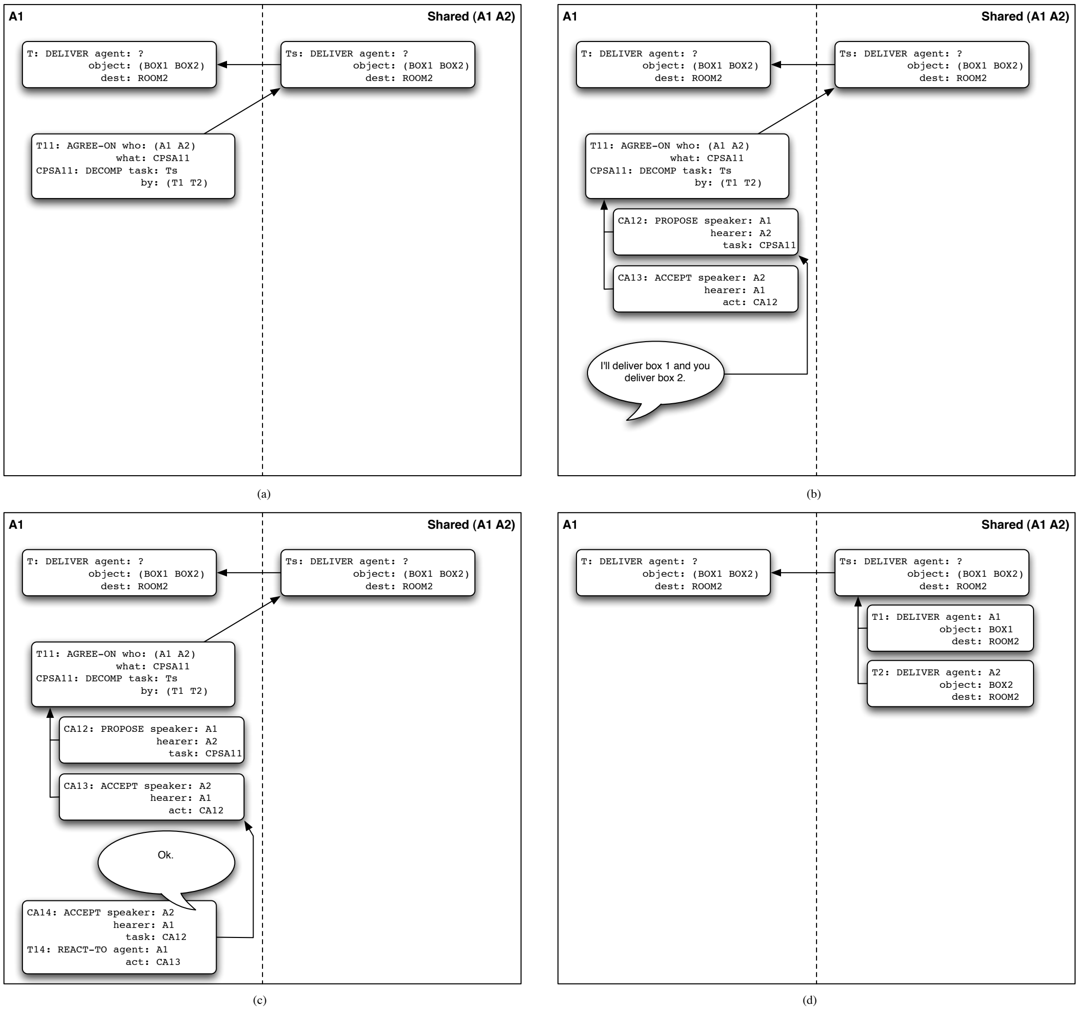

Figure 3: Collaborative planning

See Figure 3(b). Note that this follows quite naturally the exchange from the previous section (Figure 2).

Figure 3(c) show the situation after A2 replies with "*Ok.*" Again, this is the simplest case. A1's reacting to the utterance results in A2's act CA14 being matched against A1's expectation CA13. Thus CA13 is done, and so also is T11 (the agreement). In this case, the agreement leads to a new belief constructed by applying the decomposition operation in the shared task space (Figure 3(d)). Thus A1 has arrived at a similar plan for performing T as before (Figure 1), but this time both accomplishing the task and the means of accomplishing it are joint commitments.

Before moving on, let us very briefly consider what happens if A2 disagrees and rejects A1's proposal. Accounting for this requires that the model of the AGREE-ON task be more elaborate than we have shown thus far. Essentially, it needs to accommodate the possibility of one's proposal being rejected. In this case, the task is still completed, but it does not result in a change to the shared tasks. Thus our approach is to make the task model more complex, which simplifies intention recognition and is needed anyway to drive the agent's behavior.

Other Initiative and Intention Recognition

Thus far, our examples have all involved A1 taking the initiative to solve problems. This makes sense, since by assumption A1 has the goal of delivering the boxes (performing task T). But suppose that after A2 has agreed to making T a shared task (Figure 2), A2 seizes the initiative and suggests a means of accomplishing the task. This would be very natural in a human collaboration, for example. For simplicity, let's assume that A2 proposes the same way of proceeding by saying "*You deliver box 1 and I'll deliver box 2.*"

The basic interpretation process for A1 identifies the illocutionary (or conventional) act for the utterance (see, *e.g.*, (Hinkelman and Allen 1989)). In this case, there are in fact two possible interpretations. First, A2 might be proposing a new shared task: that the two agents work together on T1 and T2, independent of any prior shared tasks. That would be a proposal of a new shared task as in Figure 2. The other possible interpretation is that A2 is proposing accomplishing task Ts by performing T1 and T2 (that is DECOMP, as

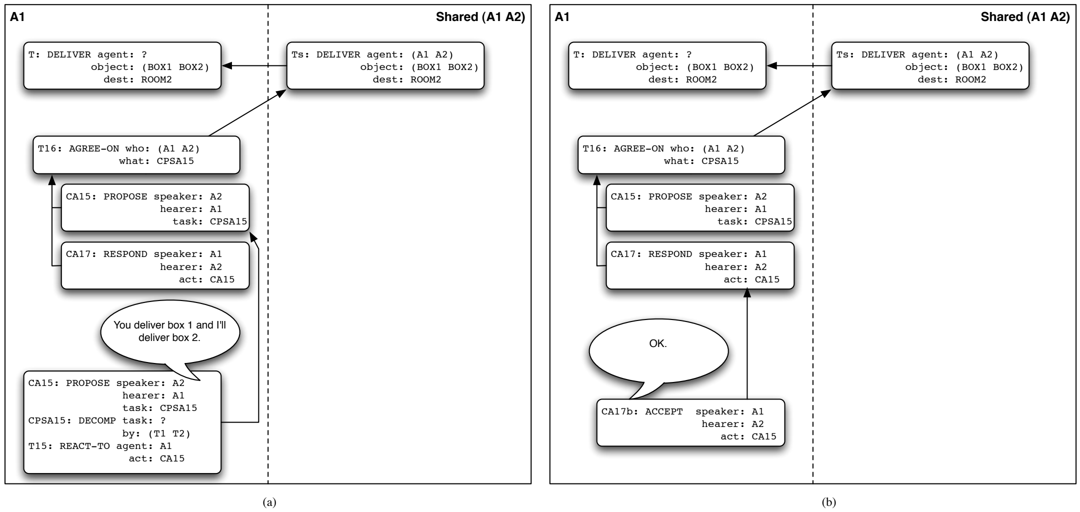

Figure 4: Other initiative and intention recognition

in Figure 3). The heuristics based on dialogue coherence described in (Allen and Litman 1990) are used to prefer the latter interpretation when it is plausible. The result is shown in Figure 4(a).

As discussed above, responding to a proposal involves either accepting or rejecting (but not ignoring) it. Thus this more elaborate model of the AGREE-ON task involves a RESPOND subtask for A1. A1 must then decide how to decompose that subtask. If it chooses to ACCEPT, the situation is as shown in Figure 4(b). The result of agreeing is to apply the DECOMP act in the shared task space, as in Figure 3(d). If A1 had instead chosen to REJECT, then the agreeing task T16 is still marked as completed, but the shared tasks are not changed. Significantly, the result of the agreement on a collaborative task is independent of which agent took the initiative to accomplish it.

Meta-level Collaboration

Finally, to illustrate the generality of our model, consider the following exchange:

A1: We need to deliver the boxes to room 2.
A2: Ok.
A1: So how should we do that?

The first two utterances are agreeing on a shared task, as in Figure 2. The third utterance is motivated by the "meta" knowledge that one way to accomplish a shared task is to first agree on the right way to perform it. In other words, agent A1 is suggesting discussing the overall problem solving strategy before getting down to specific details about the actual plan. This is especially important for complex tasks where agents need to focus their attention on different aspects of the overall task. It also allows agents to negotiate to what extent a task is shared and what aspects are left to the individual agents (*e.g.*, they agree that agent A2 will move BOX2, but leave the details entirely up to A2). Meta-talk is crucial for effective collaborative problem solving and common in human problem solving interactions.

Back to our simple example, application of this strategy by A1 yields the situation in Figure 5(a). Decomposing the meta-agreement task as a PROPOSE-ACCEPT sequence yields Figure 5(b) and A1's utterance from the exchange above. Assuming A2 accepts the proposal, the nested agreement task is established as the shared means of accomplishing the shared task (Figure 5(c)).

Many continuations are possible. A1 could take the initiative and attempt to advance the task itself, by privately planning a PROPOSE-ACCEPT sequence. For example, it could determine as before that performing T1 and T2 is the way to accomplish Ts, and propose that, as shown in Figure 5(d). Or, similarly to Figure 1, it could request that A2 perform the PROPOSE step, as shown in Figure 5(e).

Or interestingly, A1 could choose to apply the meta-level agreement operator once again, as shown in Figure 5(f). Essentially this is saying that the best way to agree on something is to agree on how to agree on it. While such an operator could be applied endlessly, this would be a bad strategy for an intelligent agent. However it is not ruled out by the model, and can be observed happening in many faculty meetings where the discussion remains at the meta-level and no actual decisions are ever made. We therefore believe we may have developed the the first formal account of unproductive meetings!

Concluding Remarks

We have developed a practical agent model for collaborative systems that supports a wide range of behavior, including private planning, collaborative planning, and the planning and use of meta-acts about the collaborative planning process itself. By practical, we mean that the agent's behavior is driven by explicit task models, rather than first principles reasoning about beliefs and intentions. This greatly simplifies both intention recognition and the planning of conversational acts.

We have developed this model by generalizing the intentional state of the agent while retaining the same basic

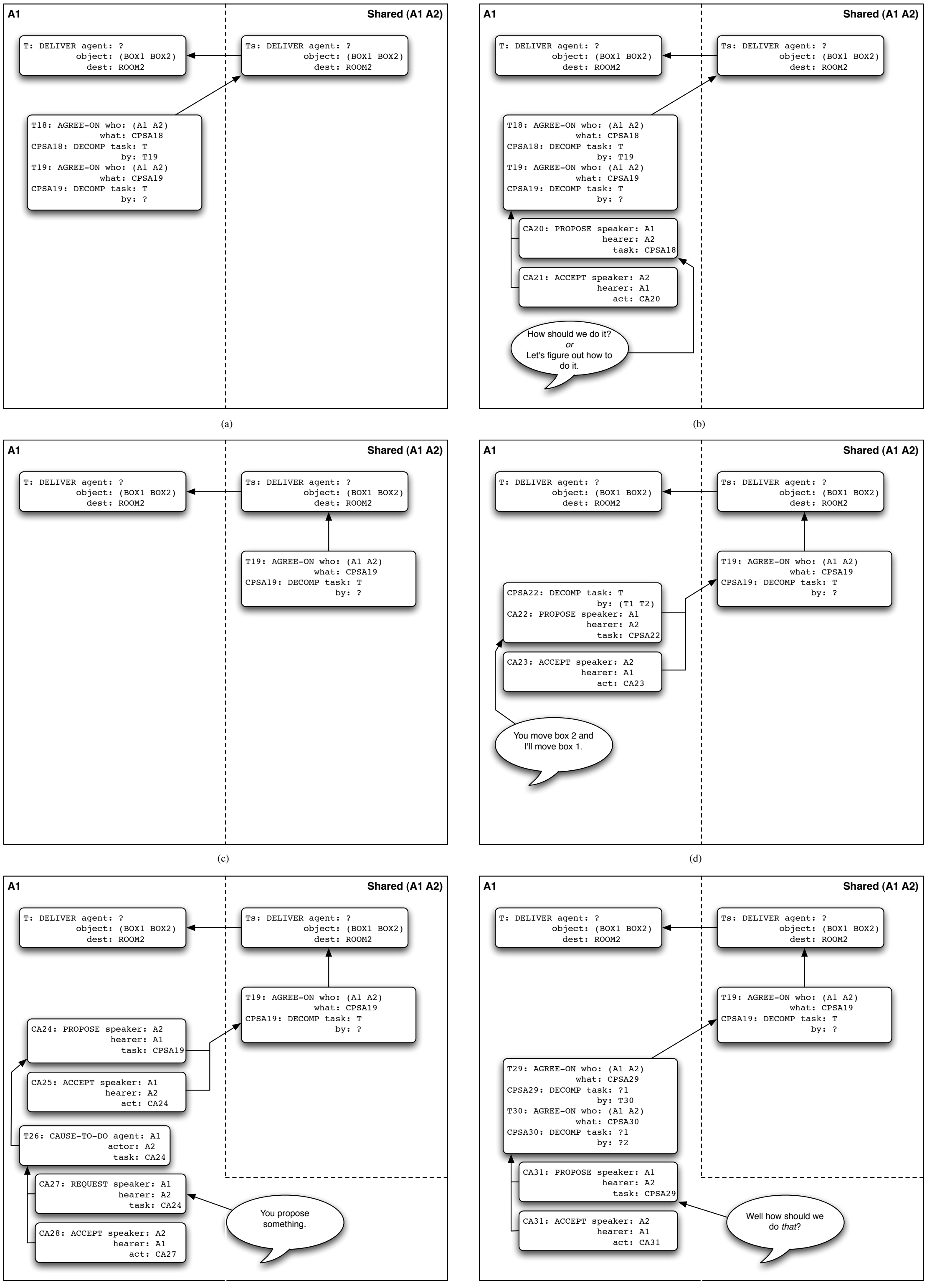

Figure 5: Meta-level collaboration

perceive-reason-act cycle used for private behavior. We consider this an essential property, as individual agents can only do individual reasoning and acting. By making all collaborative actions explicit in the model, we allow for meta-collaborative actions and open the door for learning new better ways of collaborating and planning.

Furthermore we have proposed what we think is the minimal amount of additional mechanism over the individual agent model, primarily the ability to represent shared task models and to perform intention recognition. All of the ideas, however, are justified and developed in the rich prior literature formalizing speech act planning, shared plans, and joint intentions. Our contribution here is a practical system for collaborative planning and communication cast within a fairly typical cognitive architecture. We are implementing the model in the context of practical dialogue systems for mixed-initiative decision support.

Acknowledgements

This work is supported by NSF award no. IIS-101648, DARPA OBTW award W911NF-11-2-0015-01, and ONR award no. N00014-11-1-0417.

References

Allen, J. F., and Litman, D. J. 1990. Discourse processing and commonsense plans. In Cohen, P. R.; Morgan, J.; and Pollack, M. E., eds., *Intentions and Communication*. MIT Press. 365–388.

Allen, J. F., and Perrault, C. R. 1980. Analyzing intention in utterances. *Artificial Intelligence* 15(3):143–178.

Allen, J.; Chambers, N.; Ferguson, G.; Galescu, L.; Jung, H.; Swift, M.; and Taysom, W. 2007. PLOW: A collaborative task learning agent. In *Proceedings of the Twenty-Second Conference on Artificial Intelligence (AAAI-07)*, 1514–1519.

Allen, J.; Blaylock, N.; and Ferguson, G. 2002. A problem-solving model for collaborative agents. In *Proceedings of the Conference on Autonomous Agents and Multi-Agent Systems (AAMAS-02)*, 774–781.

Bohus, D., and Rudnicky, A. 2009. The RavenClaw dialog management framework: Architecture and systems. *Computer Speech & Language* 23(3):332–361.

Clark, H., and Schaefer, E. F. 1987. Collaborating on contributions to conversations. *Language and Cognitive Processes* 19–41.

Clark, H. 1996. *Using Language*. Cambridge University Press.

Cohen, P. R., and Levesque, H. J. 1990. Intention is choice with commitment. *Artificial Intelligence* 42(2–3):213–361.

Cohen, P. R., and Levesque, H. J. 1991. Teamwork. *Nous* 25(4):487–512.

Cohen, P. R., and Perrault, C. R. 1979. Elements of a plan-based theory of speech acts. *Cognitive Science* 3:177–212. Also in *Readings in Artifical Intelligence*, B. L. Webber and N. J. Nilsson (eds.), 1981, pp. 478–495.

Ferguson, G., and Allen, J. F. 2006. Mixed-initiative dialogue systems for collaborative problem-solving. *AI Magazine* 28(2):23–32.

Georgeff, M. P., and Lansky, A. L. 1987. Reactive reasoning and planning. In *Proceedings of the Sixth National Conference on Artificial Intelligence (AAAI-87)*, 667–682. Also in *Readings in Planning*, J. Allen, J. Hendler, and A. Tate (eds.), Morgan Kaufmann, 1990, pp. 729–734.

Grosz, B. J., and Kraus, S. 1996. Collaborative plans for complex group action. *Artificial Intelligence* 86(2):269–357.

Hinkelman, E. A., and Allen, J. F. 1989. Two constraints on speech act ambiguity. In *Proceedings of the Twenty-Seventh Annual Meeting of the Association for Computational Linguistics (ACL-89)*, 212–219.

Ingrand, F. F.; Georgeff, M. P.; and Rao, A. S. 1992. An architecture for real-time reasoning and control. *IEEE Expert* 7(6):34–44.

Langley, P.; Choi, D.; and Rogers, S. 2009. Acquisition of hierarchical reactive skills in a unified cognitive architecture. *Cognitive Systems Research* 10:316–332.

Levesque, H. J.; Cohen, P. R.; and Nunes, J. H. T. 1990. On acting together. In *Proceedings of the Eighth National Conference on Artificial Intelligence (AAAI-90)*, 94–99.

Mailler, R., and Lesser, V. 2004. Solving distributed constraint optimization problems using cooperative mediation. In *Proceedings of AAMAS-2004*.

Modi, P.; Shen, W.; Tambe, M.; and Yokoo, M. 2005. ADOPT: Asynchronous distributed constraint optimization with quality guarantees. *Artificial Intelligence* 161:149–180.

Perrault, C. R.; Allen, J. F.; and Cohen, P. R. 1978. Speech acts as a basis for understanding dialogue coherence. In *Theoretical Issues in Natural Language Processing*, 125–132.

Rich, C., and Sidner, C. L. 1997. COLLAGEN: When agents collaborate with people. In *First International Conference on Autonomous Agents*, 284–291.

Roth, M.; Simmons, R.; and Veloso, M. 2005. Reasoning about joint beliefs for execution-time communication decisions. In *Proceedings of AAMAS-2005*.

Subramanian, R. A.; Kumar, S.; and Cohen, P. R. 2006. Integrating joint intention theory, belief reasoning, and communicative action for generating team-oriented dialogue. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI-06)*, 1501–1506.

Tambe, M. 1997. Towards flexible teamwork. *Journal of Artificial Intelligence Research* 7:83–124.

Taylor, M. E.; Jain, M.; Kiekintveld, C.; young Kwak, J.; Yang, R.; Yin, Z.; ; and Tambe, M. 2010. Two decades of multiagent teamwork research: Past, present, and future. In *Postproceedings of Collaborative Agents REsearch and Development (CARE) 2010 Workshop*.

Vere, S., and Bickmore, T. 1990. A basic agent. *Computational Intelligence* 6(1):41–60.

Wu, F.; Zilberstein, S.; and Chen, X. 2009. Multi-agent online planning with communication. In *Proceedings of ICAPS-2009*.