

Ten Challenges for Making Automation a “Team Player” in Joint Human-Agent Activity

Gary Klein
Klein Associates Inc.

David D. Woods
Cognitive Systems Engineering Laboratory

Jeffrey M. Bradshaw
Robert R. Hoffman
Paul J. Feltovich
Institute for Human and Machine Cognition

In this essay, we outline ten challenges for making automation components into effective “team players” when they interact with people in significant ways. Our analysis is based on some of the principles of human-centered computing that we have developed individually and jointly over the years, and is adapted from a more comprehensive examination of common ground and coordination (Klein, Feltovich, Bradshaw, & Woods, in press).

Requirements for Joint Activity

We define a joint activity is an extended set of actions that are carried out by an ensemble of people who are coordinating with each other (Clark, 1996; Klein, Feltovich, Bradshaw, & Woods, in press).

Four basic requirements for joint activity are that: all of the parties have to enter into a Basic Compact that they intend to work together, there must be some degree of interpredictability in their actions, they have to be directable by each other, and they have to maintain common ground. We expand on each of these requirements.

In order to carry out the joint activity, the parties effectively enter into a “*Basic Compact*”: an agreement (often tacit) to facilitate coordination, work toward shared goals, and prevent the team’s breakdown. The Basic Compact involves a commitment to some degree of goal alignment— typically this entails one or more participants relaxing some shorter-term goals in order to permit more global and long-term goals to be addressed. These longer-term goals might be shared goals (e.g., a relay team) or individual goals (e.g., highway drivers wanting to ensure their own safe journeys).

The Basic Compact is not a once-and-for-all prerequisite to be satisfied, but rather is a continuously reinforced or renewed. It includes an expectation that the parties will repair faulty mutual knowledge, beliefs, and assumptions when these are detected. Part of achieving coordination is investing in those actions that enhance the integrity of the compact as well as being sensitive to and counteracting those factors that could degrade it.

For example, remaining in a compact during a conversation is visible in the process of accepting turns, relating understandings, detecting the need for and engaging in repair, displaying a posture of interest, and the like. When these sorts of things are not happening, it may be inferred that one or more of the parties is not wholeheartedly engaged. The Basic Compact requires that if one party intends to drop out of the joint activity, he or she must signal this to the other parties. Breakdowns occur when a party abandons the team without clearly signaling its intentions to others.

While driving in traffic there may be defensible motives for drivers to escape from the Basic Compact to follow the rules of the road, as when they are responding to an emergency by rushing someone to the nearest hospital. At such times, drivers may turn on their emergency blinkers to signal to other drivers that their actions are no longer as predictable. But in most kinds of joint activity, the agreement itself was tacit and the partners depend on more subtle types of signaling to convey that they are or are not continuing in the joint activity. In a given

context, sophisticated protocols may develop to acknowledge the receipt of signals, transmit some construal of the meaning of the signal back to the sender, indicate preparation for consequent acts, and so forth.

For effective coordination to take place during the course of the joint activity, there has to be a reasonable level of *interpredictability*. In highly interdependent activities, it becomes possible to plan one's own actions (including coordination actions) only when what others will do can be accurately predicted. Skilled teams become interpredictable through shared knowledge and idiosyncratic coordination devices developed through extended experience in working together; bureaucracies with high turnover compensate for experience by substituting explicit predesigned structured procedures and expectations.

Team members also need to be *directable*. Directability refers to the capacity for deliberately assessing and modifying the actions of the other parties in a joint activity as conditions and priorities change (Christofferson & Woods, 2002). Effective coordination requires responsiveness of each participant to the influence of the others as the activity unfolds.

Finally, effective coordination requires establishing and maintaining *common ground* (Clark & Brennan, 1991). Common ground includes the pertinent knowledge, beliefs and assumptions that are shared among the involved parties. Common ground enables each party to comprehend the messages and signals that are typically central to coordination of joint actions. Team members must be alert for signs of possible erosion of common ground, and take preemptive action to forestall a potentially disastrous breakdown.

Consider an exercise in which a Brigade Commander had to act like an aide, in order to ensure that a staff member had seen a key piece of information on the large-format command post display that showed a "common operating picture". During the exercise, a critical event occurred and was entered into the display. The commander was not sure that one of his staff members had seen the change. Therefore, he made a radio call to the staff member because he felt it was important to manage the attention of his subordinate, and because the technology did not let him see if the staff member had noticed the event.

Making Automation A Team Player

Many researchers and system developers have been looking for ways to make automation systems team players (Christoffersen & Woods, 2002). A great deal of the current work to determine how to build automated systems with sophisticated team player qualities is taking place within the software and robotic agent research communities (e.g., Allen *et al.*, 2002; Bradshaw *et al.*, 2004a; Tambe *et al.*, 1999).¹ In contrast to early research that focused almost exclusively on how to make agents more autonomous, much of current agent research seeks to understand and satisfy requirements for the basic aspects of joint activity, either within multi-agent systems or as part of human-agent teamwork.

¹ Wherever we use the term "agent" in the remainder of the essay, we refer to software and robotic agents.

Given the widespread demand for increasing the effectiveness of team play for complex systems that work closely and collaboratively with people, a better understanding of the major challenges for making automation components into effective team players is important.

Research Challenge 1: *To be a team player, an agent must fulfill the requirements of a Basic Compact to engage in common grounding activities.*

A common occurrence in joint action is when an agent fails and can no longer perform its role. General-purpose agent teamwork models typically entail that each team member be notified of the failure (e.g., Cohen and Levesque, 1991).

Looking beyond current research and current machine capabilities, not only do agents need to be able to enter into a Basic Compact, they must also “understand” and accept the joint goals of the enterprise, understand and accept their roles in the collaboration and the need for maintaining common ground, and be capable of signaling if it is unable or unwilling to fully participate in the activity.

Research Challenge 2: *To be an effective team player, agents must be able to adequately model the other participants’ intents and actions vis-à-vis the state and evolution of the joint activity—e.g., are they having trouble? Are they on a standard path proceeding smoothly? What impasses have arisen? How have others adapted to disruptions to the plan?*

In the limited realm of what today’s agents can communicate and reason about among themselves, there has been some limited success in the development of theories and implementations of multi-agent cooperation not directly involving humans. The key concept in that work usually involves some notion of shared knowledge, goals, and intentions that function as the glue that binds the agents’ activities together (Cohen & Levesque, 1991). By virtue of a largely reusable explicit formal model of shared “intentions,” multiple agents attempt to manage general responsibilities and commitments to each other in a coherent fashion that facilitates recovery when unanticipated problems arise.

Addressing human-agent teamwork presents a new set of challenges and opportunities for agent researchers. No form of automation today or on the horizon is capable of entering fully into the rich forms of Basic Compact that are used among people.

Research Challenge 3: *Human-agent team members must be interpretable.*

To be a team player, an agent—like a human—has to be reasonably predictable, and has to have a reasonable ability to predict the actions of others. Thus it should act neither capriciously nor unobservably, and it should be able to observe and correctly predict future behavior of teammates. Currently, however, the “intelligence” and autonomy of agents directly works against the confidence that people have in their predictability. Although people will rapidly confide tasks to simple deterministic mechanisms whose design is artfully made transparent, they naturally are reluctant to trust complex agents to the same degree (Bradshaw *et*

al., 2004b). Ironically, by making agents more adaptable, we may also make them less predictable. As many have pointed out, the more a system takes the initiative in adapting to the existing working style of its operator, the more reluctant operators may be to adapt their own behavior because of the confusions these adaptations might create (e.g., Klein, 2004).

Research Challenge 4: Agents must be directable.

The non-transparent complexity and inadequate directability of agents can be a formula for disaster. In response to this concern, agent researchers have increasingly focused on developing means for controlling aspects of agent autonomy in a fashion that can both be dynamically specified and humanly understood—that is directability (Christofferson & Woods, 2002; Myers & Morley, 2003). “Policies” are a means to dynamically regulate the behavior of a system without changing code or requiring the cooperation of the components being governed (Bradshaw *et al.*, 2004a, b). Through policy, people can precisely express bounds on autonomous behavior in a way that is consistent with their appraisal of an agent’s competence in a given context. Their behavior becomes more predictable in respect to the actions controlled by policy. Moreover, the ability to change policies dynamically means that poorly performing agents can be immediately brought into compliance with corrective measures.

Research Challenge 5: Agents must be able to make pertinent aspects of their status and intentions obvious to their teammates.

Classic results have shown that the highest levels of automation on the flight deck of commercial jet aircraft (Flight Management Systems or FMS) often leave commercial pilots baffled, wondering what the automation is currently doing, why it is doing that, and what it is going to do next (reviewed in Woods and Sarter, 2000). To make their actions sufficiently predictable, agents need to make their own targets, states, capacities, intentions, changes, and upcoming actions obvious to the people and other agents that supervise and coordinate with them (e.g., Feltovich *et al.*, 2004). This challenge runs counter to the advice that is sometimes given to automation developers to create systems that are barely noticed. We are asserting that people need to have a model of the machine as an agent participating in the joint activity (Norman, 1990)—what Hoffman and Woods (2004) call the mirror-mirror principle. People can often effectively use their own thought processes as a basis for inferring the way their teammates are thinking. But this self-referential heuristic is not usually effective in working with agents.

Research Challenge 6: Agents must be able to observe and interpret pertinent signals of status and intentions.

Sending signals is not enough. The agents that receive signals have to be able to interpret the signals. The ideal agent would grasp the significance of such things as pauses, rapid pacing, and public representations that help humans to mark the coordination activity. Few existing agents are intended to read the signals of their operator teammates with any degree of substantial understanding, let alone nuance. As a result, the devices are unable to recognize the stance of the operator, much less appreciate the operator’s knowledge, mental models, or goals given the evolving state of the plan in progress and the world being controlled.

Billings (1996) and Woods (2002) have voiced their skepticism about the value of such research. They argue that an inherent asymmetry in coordinative competencies between people and machines will always create difficulties for designing human-agent teams. Nevertheless, some researchers are exploring ways to stretch the performance of agents in order to reduce this asymmetry as far as possible, such as exploiting and integrating available channels of communication from the agent to the human, and conversely sensing and inferring cognitive state through a range of physiological measures of the human in real time. Similarly, a few research efforts are taking seriously the agent's need to interpret the physical environment. If they accomplish nothing more, efforts such as these can help us appreciate the difficulty of this problem.

Research Challenge 7: Agents must be able to engage in goal negotiation.

To be a team player, an entity has to be able to enter into goal negotiation, particularly when the situation changes and the team has to adapt. As required, agents need to convey their current and potential goals so that all members of a team can participate in the negotiations.

If agents are unable to readily represent, reason about, or modify their goals, they will interfere with coordination and the maintenance of common ground. Traditional planning technologies for agents typically take an autonomy-centered approach, with representations, mechanisms, and algorithms that have been designed to ingest a set of goals and produce output as if they can provide a complete plan that handles all situations. This approach is not compatible with what we know about optimal coordination in human-agent interaction.

Research Challenge 8: Planning and autonomy support technologies must enable a collaborative approach.

A collaborative autonomy approach assumes that the processes of understanding, problem solving, and task execution are necessarily incremental, subject to negotiation, and forever tentative (Bradshaw, *et al.*, 2004c). Thus, every element of an "autonomous" system will have to be designed to facilitate the kind of give-and-take that quintessentially characterizes natural and effective teamwork among groups of people.

Allen's research on a Collaboration Management Agent (CMA) is a good example (Allen *et al.*, 2002). The CMA is designed to support human-agent, human-human, and agent-agent interaction and collaboration within mixed human-robotic teams. It interacts with individual agents in order to a) maintain an overall picture of the current situation and status of the overall plan, as completely as possible based on available reports; b) detect possible failures that become more likely as the plan execution evolves and to invoke replanning; c) evaluate the viability of proposed changes to plans by agents; d) manage replanning when situations exceed the capabilities of individual agents, including recruiting more capable agents to perform the replanning; e) manage the re-tasking of agents when changes are made; f) adjust its communications to the capabilities of the agents (e.g., graphical interfaces work well for a human but wouldn't help most agents). Because the team members will be in different states depending on how much of their original plan they have executed, the CMA must support further negotiation and re-planning at runtime.

Research Challenge 9: Agents must be able to participate in the management of attention.

As a part of maintaining common ground during coordinated activity, team members direct each other's attention to the most important signals, activities and changes. They must do this in an intelligent and context-sensitive manner, so as to not overwhelm others with low-level messages containing minimal signal mixed with a great deal of distracting noise.

Relying on their mental models of each other, responsible team members expend effort to appreciate what each other needs to notice, within the context of the task and the current situation (Sarter & Woods, 2000). Automation can compensate for trouble (e.g., asymmetric lift due to wing icing), but currently does so invisibly. Crews can remain unaware of the developing trouble until the automation nears the limits of its authority or capability to compensate. As a result, the crew may take over too late or be unprepared to handle the disturbance once they take over, resulting in a bumpy transfer of control and significant control excursions. This general problem has been a part of several incident and accident scenarios.

It will push the limits of technology to get the machines to communicate as fluently as a well-coordinated human team working in an open visible environment. The automation will have to signal when it is having trouble and when it is taking extreme action or moving towards the extreme end of its range of authority. Such capabilities will require interesting relational judgments about agent activities. How does an agent tell when another team member is having trouble in performing a function, but is not yet failing to perform? How and when does an agent effectively reveal or communicate that it is moving towards a limit of capability? Adding threshold-crossing alarms is the usual answer to these questions in the design of agents. However, in practice, rigid and context-insensitive thresholds will typically be crossed too early (resulting in an agent that speaks up too often, too soon) or too late (resulting in an agent that is too silent speaking up too little). However, focusing on the basic functions of joint activity rather than machine autonomy has already produced promising successes (Ho, Nicolic & Sarter, in press).

Research Challenge 10: Controlling the costs of coordinated activity.

The Basic Compact commits people to coordinating with each other, and to incurring the costs of providing signals, improving predictability, monitoring the others' status, and so forth. All of these take time and energy. These coordination costs can easily get out of hand, and therefore the partners in a coordination transaction have to do what they reasonably can to keep coordination costs down. This is a tacit expectation—to try to achieve economy of effort. Achieving coordination requires continuing investment and hence the power of the Basic Compact—a willingness to invest energy and accommodate to others, rather than just performing alone in one's narrow scope and sub-goals. Coordination doesn't come for free; and coordination, once achieved, does not allow one to stop investing. Otherwise the coordination breaks down.

Keeping coordination costs down is partly, but only partly, a matter of good human-computer interface design. More than that, the agents must be able to actively seek to conform to

the needs of the operators, rather than requiring operators to adapt to them. Information handoff, which is a basic exchange in coordination phases involving humans and agents, depends on common ground and interpredictability. As we have seen, agents have to become more understandable and predictable, and more sensitive to the needs and knowledge of people.

Conclusions

The ten challenges we have presented can be viewed in different lights:

- They can be seen as a blueprint for the design and evaluation of intelligent systems—requirements for successful operation and the avoidance or mitigation coordination breakdowns.
- They can be viewed as cautionary tales about the ways that the technology can disrupt rather than support coordination: Simply relying on explicit procedures, such as common operating pictures, is not likely to be sufficient.
- They point to the basis for human-agent systems. All of the challenges have us walking a fine line between the two views of AI: the traditional view that the goal of AI is to create systems that emulate human capabilities versus the Human-Centered Computing goal of creating systems that extend human capabilities, enabling people to reach into contexts that matter for human purposes.

We can imagine in the future that some agents will be able to enter into some form of a basic compact with diminished capability (Bradshaw, *et al.*, 2004a). Agents may eventually be fellow team members with humans in the way a young child can be—subject to the consequences of brittle and literal-minded interpretation of language and events, of inability to appreciate or even attend effectively to key aspects of the interaction, of poor anticipation, and of insensitivity to nuance. In the meantime, we hope that the ten challenges we have outlined might be used to guide research in the design of team and organizational simulations that capture coordination breakdowns and other features of joint activity. Through further research, restricted types of Basic Compacts might be created that could be suitable for use in human-agent systems.

Acknowledgements

Klein Associates, Ohio State, and IHMC prepared this chapter through the support of the Advanced Decision Architectures Collaborative Technology Alliance, sponsored by the U.S. Army Research Laboratory under cooperative agreement DAAD19-01-2-0009.

References

Allen, J. F., & Ferguson, G. (2002). Human-machine collaborative planning. In Proceedings of the NASA Planning and Scheduling Workshop. Houston, TX.

Billings, C. E. (1996). Aviation automation: The search for a human-centered approach. Mahwah, NJ: Lawrence Erlbaum Associates.

Bradshaw, J. M., Feltovich, P., Jung, H., Kulkarni, S., Taysom, W., & Uszok, A. (2004a). Dimensions of adjustable autonomy and mixed-initiative interaction. In M. Klusch, G. Weiss, & M. Rovatsos (Eds.), *Computational Autonomy*. (in press). Berlin, Germany: Springer-Verlag.

Bradshaw, J. M., Beautement, P., Breedy, M., Bunch, L., Drakunov, S. V., Feltovich, P. J., Hoffman, R. R., Jeffers, R., Johnson, M., Kulkarni, S., Lott, J., Raj, A., Suri, N., & Uszok, A. (2004b). Making agents acceptable to people. In N. Zhong & J. Liu (Eds.), *Intelligent Technologies for Information Analysis: Advances in Agents, Data Mining, and Statistical Learning*. Berlin: Springer Verlag., pp. 355-400.

Bradshaw, J. M., Acquisti, A., Allen, J. F., Breedy, M., Bunch, L., Chambers, N., Galescu, L., Goodrich, M., Jeffers, R., Johnson, M., Jung, H., Lott, J., et al. (2004c). Teamwork-centered autonomy for extended human-agent interaction in space applications. In *Proceedings of the AAAI Spring Symposium* (pp. 136-140). Stanford, CA: The AAAI Press.

Christoffersen, K., & Woods, D. D. (2002). How to make automated systems team players. *Advances in Human Performance and Cognitive Engineering Research*, 2, 1-12.

Cohen, P. R., & Levesque, H. J. (1991). *Teamwork*. Technote 504. Menlo Park, CA: SRI International, March.

Clark, H. (1996). *Using Language*. Cambridge: Cambridge University Press.

Feltovich, P. J., Bradshaw, J. M., Jeffers, R., Suri, N., & Uszok, A. (2004). Social order and adaptability in animal and human cultures as analogues for agent communities: Toward a policy-based approach. In A. Omacini, P. Petta & J. Pitt (Eds.), *Engineering societies in the agents world IV* (Lecture Notes in Computer Science Series). Heidelberg, Germany: Springer-Verlag.

Ho, C-Y, Nikolic. M., Waters, M., and Sarter, N.B. (2004). Not now: supporting attention management by indicating the modality and urgency of pending tasks. *Human Factors*, in press.

Hoffman, R. R., & Woods, D. D. (2004). "The Theory of Complex Cognitive Systems." Report, Institute for Human and Machine Cognition, Pensacola, FL.

Klein, G. (2004). *The Power of Intuition*. New York: A Currency Book/Doubleday.

Klein, G., Feltovich, P. J., Bradshaw, J.M., & Woods, D.D. (in press). Common ground and coordination in joint activity. In W.R. Rouse & K.B. Boff (Eds.), *Organizational Simulation*. New York: Wiley.

Myers, K., & Morley, D. (2003). Directing agents. In H. Hexmoor, C. Castelfranchi & R. Falcone (Eds.), *Agent Autonomy* (pp. 143-162). Dordrecht, The Netherlands: Kluwer.

Norman, D. A. (1990). The "problem" with automation: Inappropriate feedback and interaction, not "over-automation." *Philosophical Transactions of the Royal Society of London*, 327, 585-593.

Sarter, N., & Woods, D. D. (2000). Team play with a powerful and independent agent: A full mission simulation. *Human Factors*, 42, 390-402.

Tambe, M., Shen, W., Mataric, M., Pynadath, D. V., Goldberg, D., Modi, P. J., Qiu, Z., & Salemi, B. (1999). Teamwork in cyberspace: Using TEAMCORE to make agents team-ready. In *Proceedings of the AAI Spring Symposium on Agents in Cyberspace*. Menlo Park, CA: The AAI Press.

Woods, D.D. (2002). Steering the reverberations of technology change on fields of practice: Laws that govern cognitive work. *Proceedings of the 24th Annual Meeting of the Cognitive Science Society*. URL:<http://csel.eng.ohio-state.edu/laws>

Woods, D. D. & Sarter, N. (2000). Learning from Automation Surprises and Going Sour Accidents. In N. Sarter and R. Amalberti (Eds.), *Cognitive Engineering in the Aviation Domain*, Erlbaum, Hillsdale NJ.

BIOSKETCHES

Gary Klein is Chief Scientist of Klein Associates, Inc. Contact him at Klein Associates, Inc., 1750 Commerce Center Blvd. North Fairborn, OH 45324: gary@decisionmaking.com.

David D. Woods is Professor in the Institute for Ergonomics at the Ohio State University. Association. Contact him at the Cognitive Systems Engineering Laboratory. 210 Baker Systems, 1971 Neil Avenue, Ohio State University, Columbus, OH 43210; woods.2@osu.edu.

Jeffrey M. Bradshaw is a Senior Research Scientist at the Institute for Human and Machine Cognition. Contact him at the Institute for Human and Machine Cognition (IHMC), 40 S. Alcaniz Street, Pensacola, FL 32502; jbradshaw@ihmc.us.

Paul J. Feltovich is a Research Scientist at the Institute for Human and Machine Cognition, Pensacola, FL. Contact him at the Institute for Human and Machine Cognition (IHMC), 40 S. Alcaniz Street, Pensacola, FL 32502; pfeltovich@ihmc.us