Autonomy and Interdependence in Human-Agent-Robot Teams

Matthew Johnson, Jeffrey M. Bradshaw, and Paul J. Feltovich, *Institute for Human and Machine Cognition*

Catholijn Jonker and Birna van Riemsdijk, *Delft University of Technology*

Maarten Sierhuis, *Palo Alto Research Center*

Conventional wisdom has it that increasing the autonomy of certain classes of systems will improve their performance. For example, the United States Department of Defense Unmanned Systems Roadmap states, "The Department will pursue greater autonomy in order to improve the ability of unmanned systems to operate independently, either individually or collaboratively, to execute complex missions in a dynamic environment."[1] In the context of a report on a Gulf oil spill, a recent IEEE article suggested that "automation techniques will improve not only the time that it takes to do these tasks but also the quality of the results."[2]

General conclusions of this sort can be misleading for various reasons. One is that in a complex joint activity involving mixed teams of humans, software agents, and robots, increased autonomy can eventually lead to degraded performance when the conditions that enable effective management of interdependence among team members are neglected.

Effective interdependence management will become increasingly important in the coming years. The sophisticated robots envisioned for the future will be increasingly collaborative in nature, not merely doing things for people, but also working with people and intelligent systems. Although continuing research is needed to make agents and robots more independent when unsupervised activity is desirable or necessary—to give them autonomy, in other words—they must also be more capable of sophisticated interdependent joint activity (coactivity) when that is required. Human-agent-robot systems must support not only the fluid orchestration of task handoffs among different people and machines, but also joint participation on shared tasks requiring continuous and close interaction. Because the capabilities for coactivity interact with autonomy algorithms at a deep level, system design must incorporate them from the beginning.

1541-1672/12/$31.00 © 2012 IEEE
Published by the IEEE Computer Society

Related Work on Interdependence

In a previous article, we described the problems of what we refer to as "autonomy-centered approaches."[1] We concluded, "Even when self-directedness and self-sufficiency are reliable, matched appropriately to each other, and sufficient for the performance of the robot's individual tasks, human-robot teams engaged in consequential joint activity frequently encounter the potentially debilitating problem of opacity, meaning the inability for team members to maintain sufficient awareness of the state and actions of others to maintain effective team performance."

The literature contains many examples supporting this conclusion. Kristen Stubbs and her colleagues recently noted the lack of transparency as a problem in human-robot interaction.[2] More generally, Donald Norman identified the issue more than two decades ago, calling it "silent automation;"[3] David Woods and Nadine Sarter later used the term "automation surprises."[4] We talk about "opacity," but it's important to recognize that the challenges go far beyond simply not being able to see needed information. They can also involve predictability, directability, or other challenges that must be addressed in order to turn autonomous systems into team players.[5]

References

1. M. Johnson et al., "The Fundamental Principle of Coactive Design: Interdependence Must Shape Autonomy," *Coordination, Organizations, Institutions, and Norms in Agent Systems VI*, M. De Vos et al., eds., Springer, 2011, pp. 172–191.
2. K. Stubbs, P. Hinds, and D. Wettergreen, "Autonomy and Common Ground in Human-Robot Interaction: A Field Study," *IEEE Intelligent Systems*, vol. 22, no. 2, 2007, pp. 42–50.
3. D.A. Norman, "The 'Problem' of Automation: Inappropriate Feedback and Interaction, Not 'Over-Automation,'" *Human Factors In Hazardous Situations*, D.E. Broadbent, A. Baddeley, and J.T. Reason, eds., Oxford University Press, 1990, pp. 585–593.
4. D.D. Woods and N.B. Sarter, "Automation Surprises," *Handbook of Human Factors & Ergonomics,* G. Salvendy, ed., John Wiley & Sons, 1997, pp. 1926–1943.
5. G. Klein et al., "Ten Challenges for Making Automation a 'Team Player' in Joint Human-Agent Activity," *IEEE Intelligent Systems*, vol. 19, no. 6, 2004, pp. 91–95.

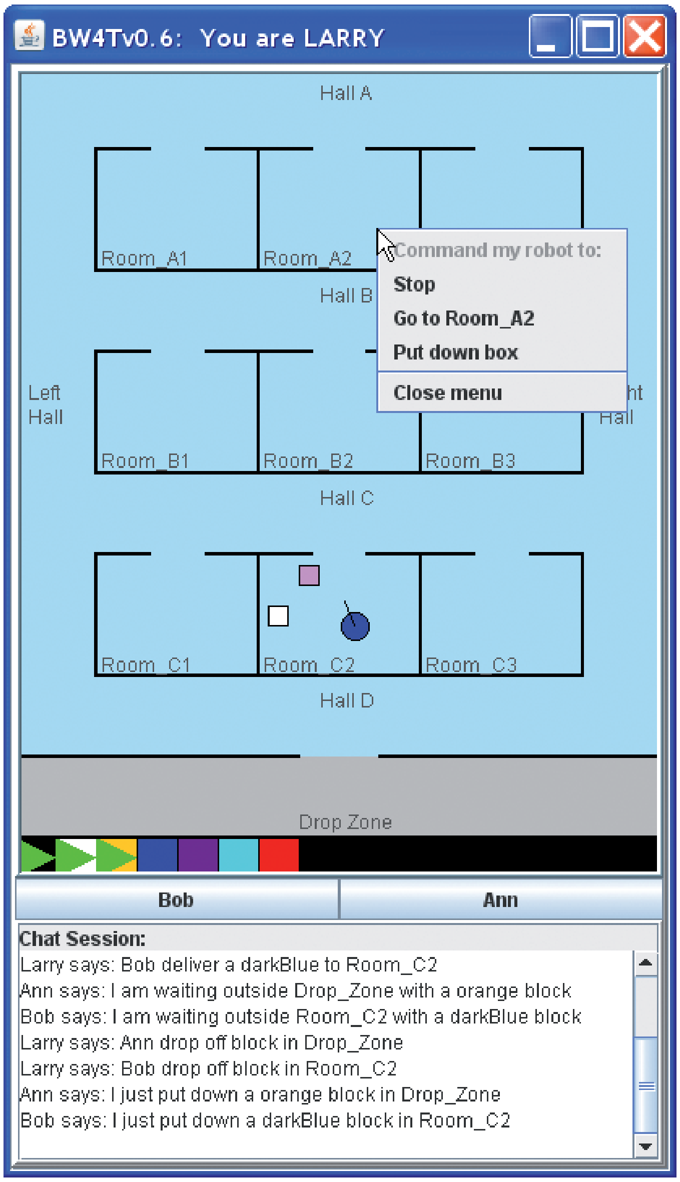
Based on this premise, our long-range goal is to develop a prescriptive methodology to guide the design and analysis of human-agent-robot systems. We intend to formulate the methodology in light of the essential role of interdependence in joint human-agent-robot activity. In this article, we examine the results of an experiment exploring how changes in autonomy can affect various dimensions of performance when interdependence is neglected. Although our experimental results stem from a simple task performed in a simulation environment, both the literature on human teamwork and our experience in a variety of human-agent-robot teamwork experiments and field exercises give us reason to believe that these results eventually can be generalized.

The Experiment

The domain for our experiment is Blocks World for Teams (BW4T), a simulation environment similar in spirit to Terry Winograd's classic AI planning problem Blocks World.[3] The goal in BW4T is to "stack" colored blocks in a particular order. The task environment comprises nine rooms containing a random assortment of blocks, plus a goal area for dropping them off (see Figure 1). Each player controls an avatar, which the player can move between rooms to pick up and drop off blocks. For this experiment, each team had two players, a human and a software agent; humans control their own avatars and command their agent partners through an appropriate interface. The two players work toward the shared team goal, which is to drop the blocks off in a specified order. Players are limited in their awareness of the situation: they can't see each other, and they can only see the blocks in their current room.

Our goal was to demonstrate that in human-agent-robot systems engaged in joint activity, increasing autonomy without addressing interdependence may lead to suboptimal performance. We attempted to eliminate failure due to overly trusting automation by ensuring that the agent players never made mistakes and that they exhibited reasonably intelligent behavior. We also didn't want the human players to manage an agent capable of completing the mission autonomously at a low level, akin to teleoperation, so we attempted to ensure the human and the agent could interact at a relatively high level of abstraction. To this end, we provided an interface appropriate to agents' capabilities. Figure 2a illustrates these elements of our experimental design.

Figure 2b shows the general trends we expected to find in our results. We anticipated that the management burden the agent player imposed on the human player would decrease as agent autonomy increased—no surprise, given that reduction in human workload is both the common expectation and the major motivation for automation. However, we also anticipated that without support for managing interdependence issues, the opacity of the work system to task participants would grow with increasing autonomy. Given these competing factors of burden and opacity, we expected to find an inflection point in team performance where the benefits of increasing autonomy eventually would be completely offset by the negative side effects of opacity. In other words, we predicted that the highest level of autonomy would not produce the highest level of team performance (see Figure 2c).



Chat Session:
Larry says: Bob deliver a darkBlue to Room_C2
Ann says: I am waiting outside Drop_Zone with a orange block
Bob says: I am waiting outside Room_C2 with a darkBlue block
Larry says: Ann drop off block in Drop_Zone
Larry says: Bob drop off block in Room_C2
Ann says: I just put down a orange block in Drop_Zone
Bob says: I just put down a darkBlue block in Room_C2

Figure 1. A sample Blocks World for Teams (BW4T) interface. The player's avatar (the blue dot) picks up boxes from the rooms and delivers them to the Drop Zone.

Defining the Agent Teammate

The algorithm we chose as the basis for the agent behavior reflects the most common approach we observed for human players of the game, because we thought it would be easily understandable and predictable for most human players. The left side of Figure 3 shows the algorithmic solution, which divides the main goal (a color sequence) into several subgoals (individual colors). To achieve any given subgoal, one simply finds the block of the appropriate color and delivers it. The tasks don't need to be performed in sequence or by the same player: a player could first find all the blocks and then deliver them, or one player could find a block and the other could deliver it. The overall task comprises several *find* tasks and several *deliver* tasks, which themselves include some decision and action primitives, such as going to a room, entering the room, going to a block, picking up a block, and putting down a block. The two main decisions are whether to look for a block or to deliver one, and which room to enter to look for a block. We designed the agent player to perform its task "perfectly," meaning it will perform any assigned task efficiently and will make rational decisions based on a complete and accurate recollection of where it has been and what it has seen. It will also report when a task is completed. To be consistent, it reports only the completion status and doesn't provide any additional information.

Defining the Autonomy Treatments

To compare the effects of changing autonomy, we defined different *autonomy treatments* (experimental conditions). Additionally, we needed some way to rank the treatments in terms of their relative degree of autonomy. For this purpose we applied Thomas Sheridan and William Verplank's concept of levels of autonomy[4] and Daniel Olsen and Michael Goodrich's neglect tolerance metric.[5] Neglect tolerance is based on the amount of time a human can ignore a given robot performing a given task before the robot becomes unproductive.

The vertical black lines in Figure 3 indicate the portion of the algorithm that the agent player performed autonomously. During those sections, the agent functioned at the highest level of autonomy, performing on its own everything necessary to complete the task specified. Longer lines cover more sections of the algorithm; thus, in general they entail more autonomy. Outside the line, the agent functioned at the lowest level of autonomy and completely relied on the human for all decisions and actions. The human teammate always initiated the behavior associated with each band. The neglect tolerance correlates to the length of the line, although the line length doesn't directly correspond to length of time, because some tasks take longer than others.

Treatment 1 required the human player to direct the agent player using only the action primitives. Consequently, the bands in treatment 1

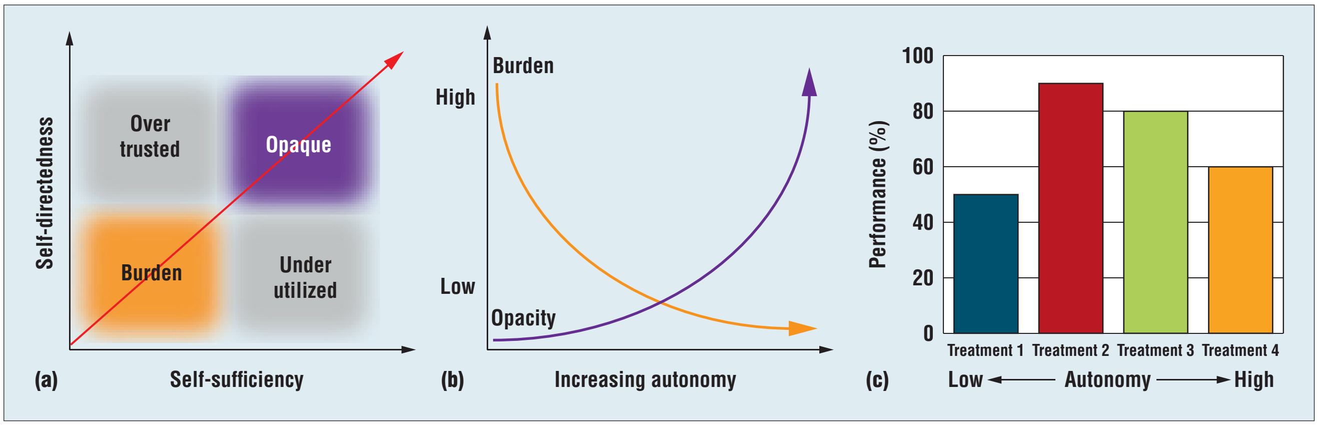


Figure 2. Experimental design and expectations: (a) balancing excessive trust in automation against underutilization of agents' autonomy, (b) expected effects of increasing autonomy on the burden of managing the agent and the opacity of the agent to other task participant, and (c) expected performance as autonomy increases.

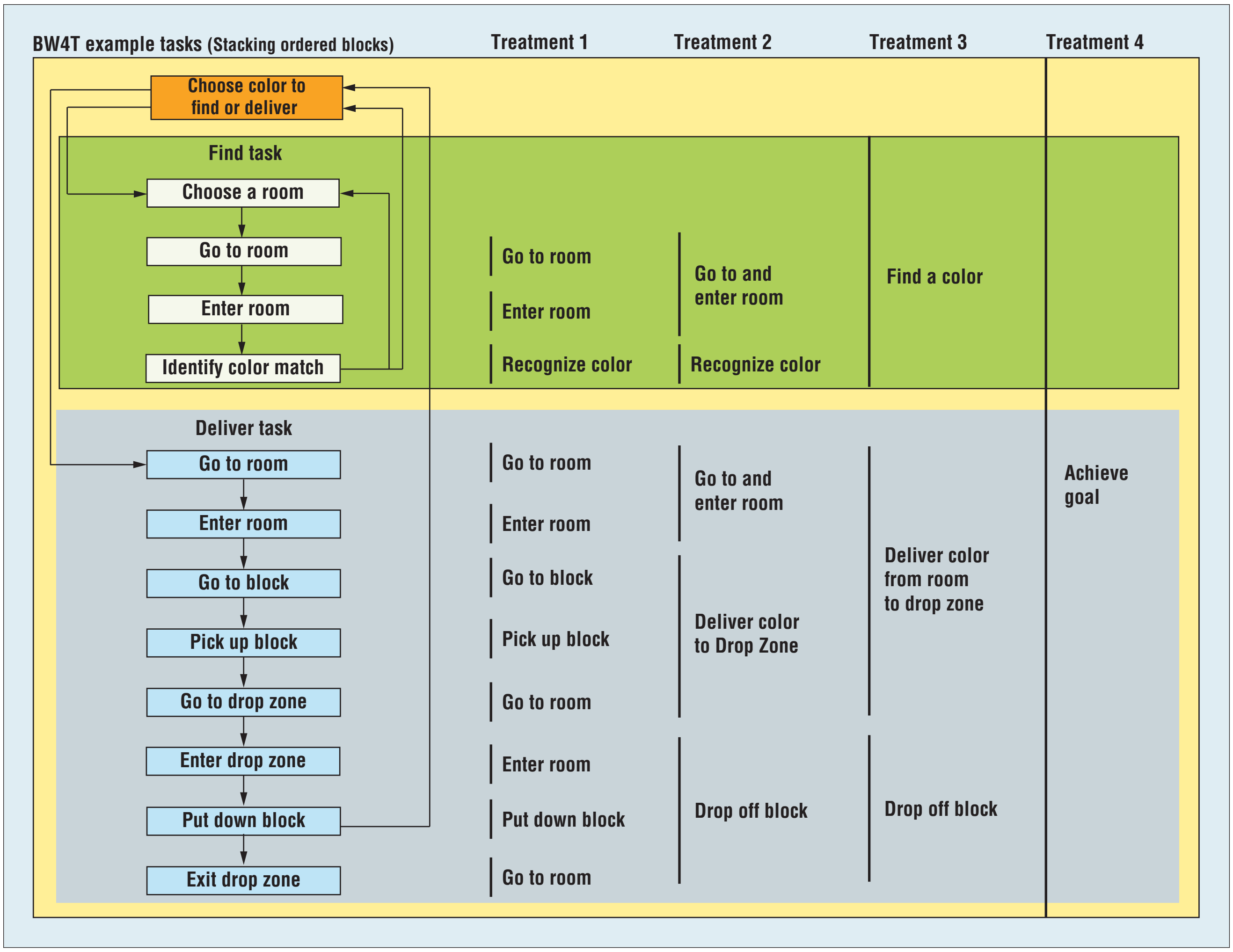


Figure 3. Defining autonomy treatments for BW4T. The left column outlines the algorithm used by the agents, and the other columns use vertical black lines to indicate the portions of the algorithm supported by autonomy for each of the four treatments.

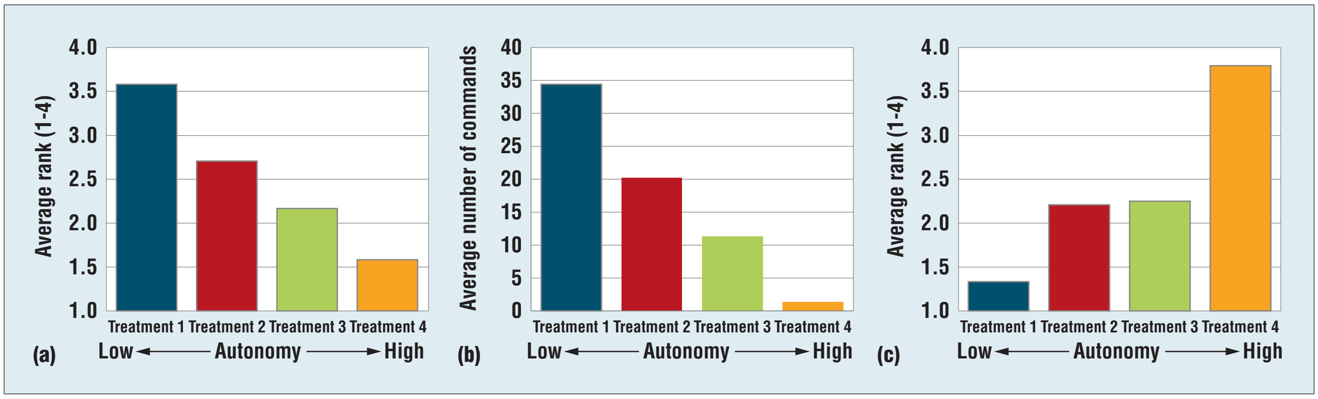


Figure 4. Relationship between management burden and autonomy: (a) subject ranking of agent management workload as autonomy increased, (b) the average number of commands required as autonomy increased, and (c) the average subjective rankings of awareness as autonomy increased.

are the shortest, the agents required more direction from their human teammate, and they had the lowest neglect tolerance.

Treatment 2 combined several action primitives into a single action. For example, with a single command the human player could now order the agent to go to and enter a room. To inhibit underutilization, we restricted the command set to the new "higher-level" commands. We only combined action primitives, so the autonomy scale doesn't provide much guidance, but it's clear that agent neglect tolerance increases, and so this treatment provides the agent more autonomy than the first.

Treatment 3 extended treatment 2 by adding the ability to command the agent to find a color or to deliver a color. This command delegated the decision on where to search to the agent, who had to provide its own search algorithm and only reported when it found a color. We implemented the command as a nearest-unsearched-room algorithm, which was the most common approach human players used in our observations. Again, we restricted the human player to the commands listed. Consistent with Sheridan and Verplank's specification for levels of autonomy, this treatment provides a higher level of autonomy than the previous one, because the agent can now make its own decision on how to achieve the find task. The level of neglect tolerance is also higher.

Treatment 4 was identical to treatment 3 but also let the agent choose whether to look for a block or deliver a block. The only required command by the human was to tell the agent to achieve the goal. This let the agent complete the entire task without any assistance from the human—in other words, it operated at the highest level of autonomy and with an infinite tolerance for neglect. The agent "decides everything, acts autonomously, ignoring the human."[6]

We intentionally left out any support for managing interdependence except for communicating task completion status. There was neither communication about world state nor coordination of task activity. Although this might seem extreme in this simple domain with obvious coordination needs, it's not an unrealistic scenario, given the prevalence of similarly opaque systems.[7–9]

Experimental Design

For the experiment, we selected 24 participants (17 male and 7 female) ranging in age from 19 to 39 from the student population at the Delft University of Technology. We used a completely randomized block design based on the autonomy treatment, with each participant performing each treatment once. We cross-classified the data by $k = 4$ autonomy treatments and $b = 24$ blocks (one for each participant). We gave all the participants a demographic survey and trained them on the game until they demonstrated proficiency at a simplified version of the task. Next they performed a series of trials, one for each treatment. The participants filled out a brief survey at the end of the experiment, evaluating team burden, opacity, performance, and preference in each treatment.

Results

Our results include quantitative numeric data as well as subjective ranking data. For the former, we use standard approaches for normal data. For the ranked data, we used the nonparametric Friedman test. Based on our design, and using the $\alpha = 0.05$ level of significance, the critical value is $\chi^2(0.95, 3) = 7.815$.

Assessing Burden

We had predicted a decrease in agent management burden as autonomy increased. We asked the participants to rank how demanding it was to work with the agent in each condition, on a scale of 1 (least demanding) to 4 (most demanding). The results, shown in Figure 4a, indicate a very clear ($\chi^2(0.95, 3) = 34.225$) decrease in burden as autonomy increased. As a

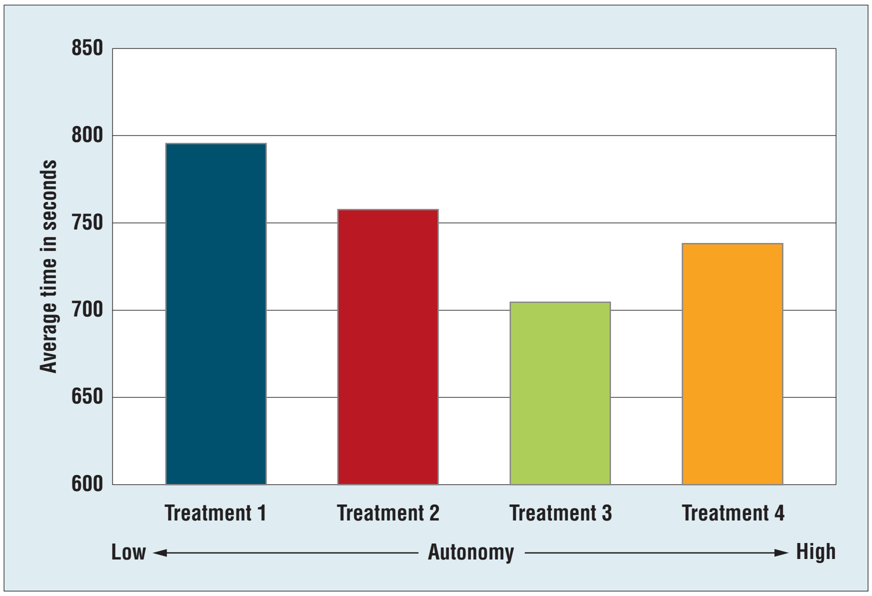


Figure 5. Time-to-completion as autonomy increases across treatments. This measure is the total time required to deliver all blocks in order.

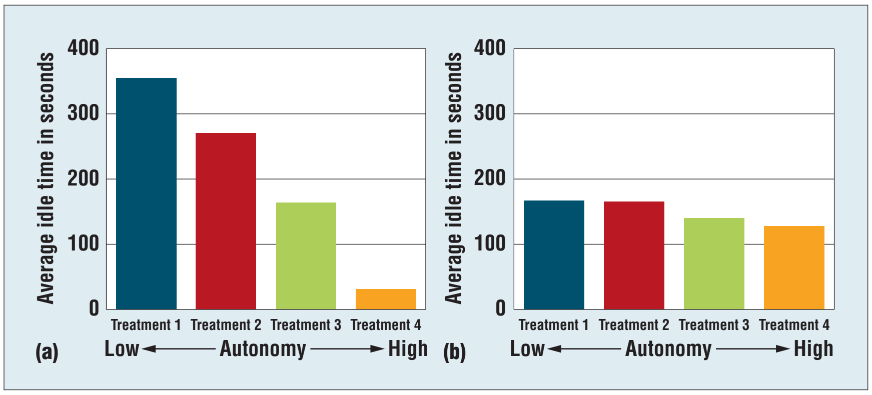


Figure 6. Average player idle time across treatment conditions: (a) agent player idle time and (b) human player idle time. The robot idle time decreases as autonomy increases, but this doesn't necessarily equate to increased effectiveness.

second, independent measure of burden, we also counted the number of commands the human player had to give to the agent teammate in each condition. Figure 4b shows the results, which correlate with the participants' subjective assessment.

Assessing Opacity

We had also predicted an increased subject perception of opacity with increasing autonomy, as revealed in participants' reports of more difficulty understanding what was happening and anticipating the agent's behavior. In an exit survey, we asked participants to rank their ongoing sense of awareness of current and future agent actions on a scale of 1 (most aware) to 4 (least aware). The results in Figure 4c show opacity increasing with increasing autonomy, as predicted ($\chi^2(.95, 3) = 49.700$). This confirms our prediction and validates our general expectations (see Figure 1b).

Quantitative Performance Assessment

We performed three quantitative performance assessments: time to complete task, idle time, and error rate.

Time to complete task. The simplest performance metric is time to completion (delivering all the required blocks in the requested order). Figure 5 shows the results, which appear promising at first glance. We can clearly see the inflection point where performance begins to degrade rather than improve as autonomy increases, consistent with the prediction of Figure 2c. The differences, however, were not statistically significant ($p = 0.20$). We believe that variability in completion time from run to run (approximately 160 seconds) was larger than the penalties from errors (such as 30 seconds of redundant activity). Nevertheless, for 83 percent of the participants, the highest-autonomy condition did not result in the lowest time.

Idle time. Another important performance measure is idle time (or wait time).[10] In BW4T, the agent player is in nearly constant motion once its human teammate assigns a task. Any idle time, such as time spent waiting for the next command, indicates an inefficient use of the agent. Figure 6a shows a clear and significant decrease in idle time from treatments 1 to 4. On the surface, this could be taken as a sign of more effective use of the agent player, suggesting improved performance. However, the time-to-completion results don't support that conclusion. Furthermore, the amount of work done—the number of rooms entered and the number of boxes delivered—is fairly consistent across treatments.

The human's idle time, shown in Figure 6b, indirectly relates to interaction efficiency, since the person's own avatar may stop while the human gives a task to the agent partner.[10] There is only a slight decrease as the agent's autonomy increases, nowhere near as

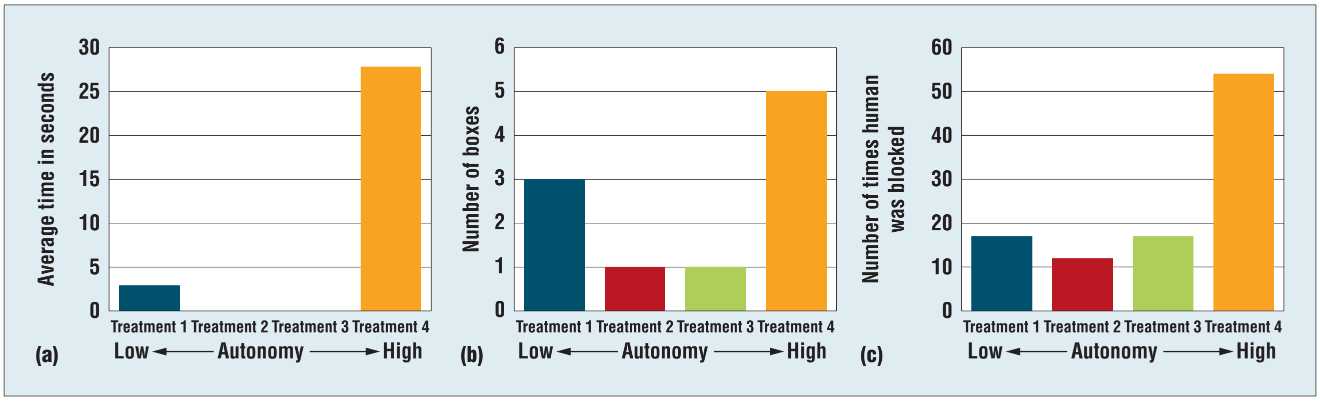


Figure 7. Error rates across treatment conditions: (a) average time players spent holding the same color box, (b) the number of lost boxes, and (c) the number of times the agent player blocked its human teammate from entering a room.

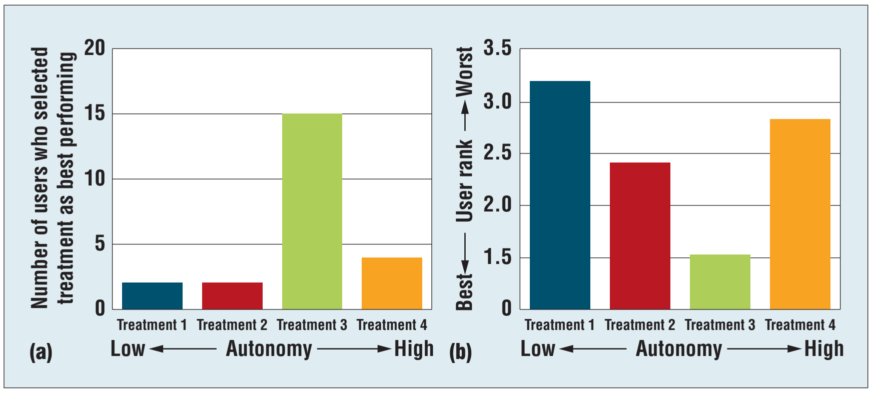


Figure 8. User impressions of the experimental treatments: (a) user assessment of performance and (b) user preference. The users neither preferred the highest-autonomy scenario nor assessed it as performing best.

large as the change seen for the agent player. This could be due to an effective interface, but it also can be due to the ability to multitask and complete interactions concurrent with motion. The interesting take-away from this result is that keeping the agent busy doesn't improve performance.

Error rate. For some kinds of tasks, error rate can be a good performance comparison. We measured this in three ways.

First was the amount of time that both players spent holding the same color block (see Figure 7a). Because the blocks the players were supposed to move were different colors, holding the same color block represented redundant or inefficient activity. This type of error, for the most part, only occurred in treatment 4 and is a side effect of the opacity of the highest-autonomy condition. These results are no surprise, because this is the only treatment in which the agent player could make its own decision about which block to pick up, but they do emphasize that functional differences matter when automating tasks.[11]

Second was the number of boxes lost (dropped in the hallway or placed in the drop zone in the wrong order). BW4T is very simple, and the human players didn't make many mistakes (and the agent players didn't make any, because they were programmed to perform perfectly). Of the 10 lost boxes, 50 percent were in treatment 4 and 30 percent in treatment 1 (see Figure 7b). The losses in treatment 1 were most likely due to the high workload imposed by the minimal autonomy. However, treatment 4 doesn't have the workload challenges of treatment 1—in fact, it was clearly ranked as the least burdensome—so why would it have the highest occurrence of errors? We believe it's a side effect of the high opacity of the highest-autonomy condition.

Our third measure of error was the number of times the agent player blocked the human player from entering a room. This is an indirect measure, because it is possible that the most efficient act would be to wait outside a blocked door, but in general it indicates poor coordination. As Figure 7c shows, the agent blocked the human player most often (by far) in treatment 4, indicating significantly more coordination breakdowns than in any other treatment.

Subjective Performance Assessment

We also assessed performance according to the participants' subjective impressions of each treatment.

Performance assessment. We asked the participants to identify which team they felt performed best. Treatment 3 was the clear winner, with 63 percent of participants selecting it (see Figure 8a). Only 17 percent chose treatment 4.

THE AUTHORS

Matthew Johnson is a research associate at the Institute for Human and Machine Cognition. Contact him at mjohnson@ihmc.us.

Jeffrey M. Bradshaw is a senior research scientist at the Institute for Human and Machine Cognition. Contact him at jbradshaw@ihmc.us.

Paul J. Feltovich is a research scientist at the Institute for Human and Machine Cognition. Contact him at pfeltovich@ihmc.us.

Catholijn Jonker is the head of the Man Machine Interaction Group of the Department of Mediametrics at the Delft University of Technology. Contact her at C.M.Jonker@tudelft.nl.

Birna van Riemsdijk is an assistant professor in the Man Machine Interaction Group of the Department of Mediametrics at the Delft University of Technology. Contact her at m.bvanriemsdijk@tudelft.nl.

Maarten Sierhuis is an area manager in the Intelligent Systems Laboratory at the Palo Alto Research Center. Contact him at Maarten.Sierhuis@parc.com.

User preference. Human acceptance is an important component of overall system performance in tasks like ours. We asked participants to rank the agents according to which one they would like to play with again, on a scale of 1 (most like to) to 4 (least like to).

Figure 8b shows the results. Users preferred treatment 3 with statistical significance ($\chi^2(.95, 3) = 22.150$). This result also demonstrates the inflection point we anticipated from the increasing opacity in the system (see Figure 2c). We suspect this is because in treatment 3, the human holds the overall plan and most of the context and exercises the greatest degree of creativity. Transparency and control may be more important than autonomy, especially in light of the particulars of the autonomous task.

We asked participants for their reasons for ranking treatment 3 higher. Responses included the following statements:

- "shared information,"
- "ability to anticipate,"
- "predictable,"
- "low burden,"
- "cleverest," and
- "automatic, but still have control."

The first three reasons correlate with our predictions about opacity. The comment about low burden is interesting, because treatment 4 was objectively the least burdensome. This comment suggests that there might be other types of burden besides the manual workload of tasking the agent. The comment about treatment 3 being cleverest is also interesting, because the agent in treatment 4 is objectively the most capable. This suggests that being more independent might not necessarily lead to being viewed as more clever. The final response is also important because it relates to the broader issue of the best way to make automation a team player.[12] We focused on opacity to keep the experiment simple, but increased autonomy no doubt also affects predictability, directability, and other challenges.

The results of our initial limited evaluation support our claim that increasing autonomy does not always improve performance of the human-machine system. In the BW4T domain, this was the result of opacity in the system due to increasing autonomy without accounting for the interdependence of the players and the coordination challenges that creates. The ability to work with others becomes increasingly important as interdependence in the joint activity grows, and in complex and uncertain domains, it might be more valuable than the ability to work independently.

In our experiment, the independent activity in treatment 4 inhibited the team's ability to engage in what most people would consider "natural" coordination, resulting in a breakdown of common ground,[14] a reduction in each player's individual situation awareness, and an increase in errors. While obvious in this simple, abstract domain, the problem remains prevalent in many systems today.[7–9] Considering interdependence when designing an agent's autonomous capabilities can mitigate these effects and will enable future systems to achieve improved results.

References

1. *Unmanned Systems Roadmap, 2007–2032*; www.fas.org/irp/program/collect/usroadmap2007.pdf.
2. A. Bleicher, "The Gulf Spill's Lessons for Robotics," *IEEE Spectrum*, vol. 47, no. 8, pp. 9–11.
3. M. Johnson et al., "Joint Activity Testbed: Blocks World for Teams (BW4T)," *Proc. Eng. Societies in the Agents World X* (ESAW 09), Springer, 2009, pp. 254–256.
4. T.B. Sheridan and W. Verplank, "Human and Computer Control of Undersea Teleoperators," Man-Machine Systems Laboratory, MIT Department of Mechanical Engineering, MIT, 1978; http://www.dtic.mil/cgi-bin/GetTRDoc?Location=U2&doc=GetTRDoc.pdf&AD=ADA057655.
5. D.R. Olsen and M. Goodrich. "Metrics for Evaluating Human-Robot Interaction," *Proc. Performance Metrics for Intelligent Systems* (PerMIS 03), Citeseer, 2003, pp. 507–527.
6. R. Parasuraman, T. Sheridan, and C. Wickens, "A Model for Types and

Levels of Human Interaction with Automation Systems," *IEEE Trans. Man and Cybernetics Part A*, vol. 30, no. 3, 2000, pp. 286–297.

7. K. Stubbs, P. Hinds, and D. Wettergreen, "Autonomy and Common Ground in Human-Robot Interaction: A Field Study," *IEEE Intelligent Systems*, vol. 22, no. 2, 2007, pp. 42–50.
8. D.A. Norman, "The 'Problem' of Automation: Inappropriate Feedback and Interaction, Not 'Over-Automation,'" *Human Factors In Hazardous Situations*, D.E. Broadbent, A. Baddeley, and J.T. Reason, eds., Oxford University Press, 1990, pp. 585–593.
9. D.D. Woods and N.B. Sarter, "Automation Surprises," *Handbook of Human Factors & Ergonomics,* G. Salvendy, ed., John Wiley & Sons, 1997, pp. 1926–1943.
10. J.W. Crandall and M.L. Cummings, "Developing Performance Metrics for the Supervisory Control of Multiple Robots," *Proc. 2nd ACM/IEEE Int'l. Conf. Human-Robot Interaction* (HRI 07), ACM, 2007, pp. 33–40.
11. M. Johnson et al., "Beyond Cooperative Robotics: The Central Role of Interdependence in Coactive Design," *IEEE Intelligent Systems*, vol. 26, no. 3, 2011, pp. 81–88.
12. G. Klein et al., "Ten Challenges for Making Automation a 'Team Player' in Joint Human-Agent Activity," *IEEE Intelligent Systems*, vol. 19, no. 6, 2004, pp. 91–95.