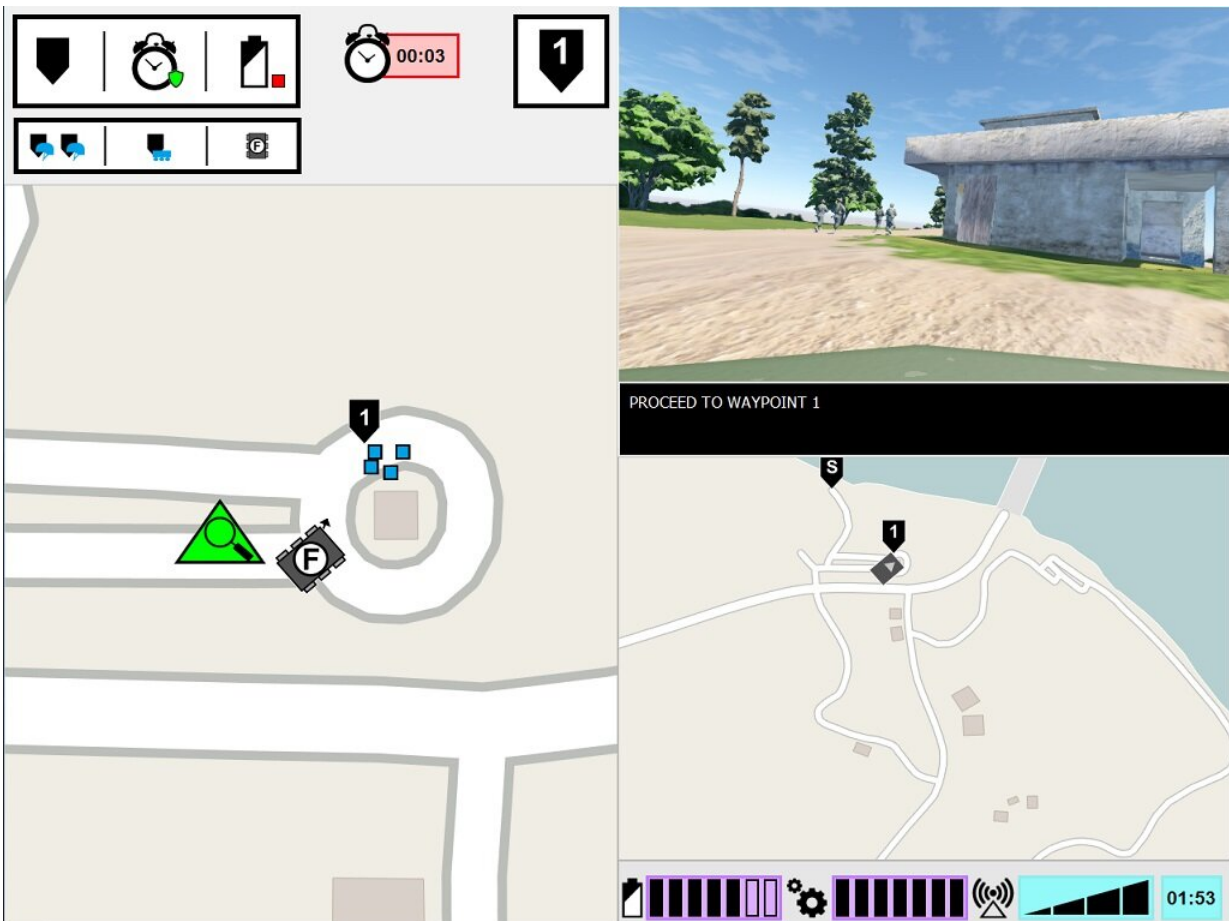# When it comes to robots, reliability may matter more than reasoning

September 23 2019

ASM experimental interface. The left-side monitor displays the lead Soldier's point of view of the task environment. The right-side monitor displays the ASM's communication interface. Credit: US Army

What does it take for a human to trust a robot? That is what Army researchers are uncovering in a new study into how humans and robots work together.

Research into human-agent teaming, or HAT, has examined how the transparency of agents—such as robots, unmanned vehicles or software agents—influences human trust, task performance, workload and perceptions of the agent. Agent transparency refers to its ability to convey to humans its intent, reasoning process and future plans.

New Army-led research finds that human confidence in robots decreases after the robot makes a mistake, even when it is transparent with its reasoning process. The paper, "Agent Transparency and Reliability in Human-Robot Interaction: The Influence on User Confidence and Perceived Reliability," has been published in the August issue of *IEEE-Transactions on Human-Machine Systems*.

To date, research has largely focused on HAT with perfectly reliable intelligent agents—meaning the agents do not make mistakes—but this is one of the few studies that has explored how agent transparency interacts with agent reliability. In this latest study, humans witnessed a robot making a mistake, and researchers focused on whether the humans perceived the robot to be less reliable, even when the human was provided insight into the robot's reasoning process.

"Understanding how the robot's behavior influences their human teammates is crucial to the development of effective human-robot teams, as well as the design of interfaces and communication methods between team members," said Dr. Julia Wright, principal investigator for this project and researcher at U.S. Army Combat Capabilities Development Command's Army Research Laboratory, also known as ARL. "This research contributes to the Army's Multi-Domain Operations efforts to ensure overmatch in artificial intelligence-enabled

capabilities. But it is also interdisciplinary, as its findings will inform the work of psychologists, roboticists, engineers, and system designers who are working toward facilitating better understanding between humans and autonomous agents in the effort to make autonomous teammates rather than simply tools.

This research was a joint effort between ARL and the University of Central Florida Institute for Simulations and Training, and is the third and final study in the Autonomous Squad Member (ASM) project, sponsored by the Office of Secretary of Defense's Autonomy Research Pilot Initiative. The ASM is a small ground robot that interacts with and communicates with an infantry squad.

Prior ASM studies investigated how a robot would communicate with a human teammate. Using the Situation awareness-based Agent Transparency model as a guide, various visualization methods to convey the agent's goals, intents, reasoning, constraints, and projected outcomes were explored and tested. An at-a-glance iconographic module was developed based on these early study findings, and then was used in subsequent studies to explore the efficacy of agent transparency in HAT.

Researchers conducted this study in a simulated environment, in which participants observed a human-agent Soldier team, which included the ASM, traversing a training course. The participants' task was to monitor the team and evaluate the robot. The Soldier-robot team encountered various events along the course and responded accordingly. While the Soldiers always responded correctly to the event, occasionally the robot misunderstood the situation, leading to incorrect actions. The amount of information the robot shared varied between trials. While the robot always explained its actions, the reasons behind its actions and the expected outcome of its actions, in some trials the robot also shared the reasoning behind its decisions, its underlying logic. Participants viewed multiple Soldier-robot teams, and their assessments of the robots were

compared.

The study found that regardless of the robot's transparency in explaining its reasoning, the robot's reliability was the ultimate determining factor in influencing the participants' projections of the robot's future reliability, trust in the robot and perceptions of the robot. That is, after participants witnessed an error, they continued to rate the robot's reliability lower, even when the robot did not make any subsequent errors. While these evaluations slowly improved over time as long as the robot committed no further errors, participants' confidence in their own assessments of the robot's reliability remained lowered throughout the remainder of the trials, when compared to participants who never saw an error. Furthermore, participants who witnessed a robot error reported lower trust in the robot, when compared to those who never witnessed a robot error.

Increasing agent transparency was found to improve participants' trust in the robot, but only when the robot was collecting or filtering information. This could indicate that sharing in-depth information may mitigate some of the effects of unreliable automation for specific tasks, Wright said. Additionally, participants rated the unreliable robot as less animate, likable, intelligent, and safe than the reliable robot.

"Earlier studies suggest that context matters in determining the usefulness of transparency information," Wright said. "We need to better understand which tasks require more in-depth understanding of the agent's reasoning, and how to discern what that depth would entail. Future research should explore ways to deliver transparency information based on the tasking requirements."

More information: Julia L. Wright et al, Agent Transparency and Reliability in Human–Robot Interaction: The Influence on User Confidence and Perceived Reliability, *IEEE Transactions on Human-*