



ARL-TR-10053 • JAN 2025



Human-in-the-Loop Feedback to Improve Robot Teammate Learning

by Julia L. Wright and Maggie Wigness

DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited.

NOTICES

Disclaimers

The research reported herein was sponsored by the U.S. Army Combat Capabilities Development Command Army Research Laboratory and was accomplished under Cooperative Agreement W911NF-19-2-0251 with the University of Central Florida. The views and conclusions contained in this report are those of the authors and should not be interpreted as representing the official policies or positions, either expressed or implied, of DEVCOM Army Research Laboratory, the U.S. Government, or the University of Central Florida. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.



Human-in-the-Loop Feedback to Improve Robot Teammate Learning

Julia L. Wright and Maggie Wigness
DEVCOM Army Research Laboratory

REPORT DOCUMENTATION PAGE

1. REPORT DATE		2. REPORT TYPE		3. DATES COVERED	
January 2025		Technical Report		START DATE January 2023	END DATE July 2024
4. TITLE AND SUBTITLE Human-in-the-Loop Feedback to Improve Robot Teammate Learning					
5a. CONTRACT NUMBER		5b. GRANT NUMBER		5c. PROGRAM ELEMENT NUMBER	
5d. PROJECT NUMBER		5e. TASK NUMBER		5f. WORK UNIT NUMBER	
6. AUTHOR(S) Julia L. Wright and Maggie Wigness					
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) DEVCOM Army Research Laboratory ATTN: FCDD-RLA-FD Orlando, FL 32826				8. PERFORMING ORGANIZATION REPORT NUMBER ARL-TR-10053	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)		10. SPONSOR/MONITOR'S ACRONYM(S)		11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION/AVAILABILITY STATEMENT DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited.					
13. SUPPLEMENTARY NOTES ORCIDs: Julia L. Wright, 0000-0003-3026-1538; Maggie Wigness, 0000-0003-1707-8106					
14. ABSTRACT Participants oversaw a robot autonomously navigating to four waypoints in a specified order. The robot initially preferred to traverse over grass, and participants were instructed to train it to drive on the road. Each participant's task was to observe the robot's actions, intervene when its behavior was undesirable, and demonstrate the desired behavior by teleoperating the robot. Participants determined when to intervene and how long to demonstrate the desired behavior. After each demonstration, the robot updated its behavior model to better mimic the demonstrated actions and then continued on using the newly updated model. Each participant completed three trials under different transparency conditions. The within-subjects analysis showed that demonstrations effectively changed the robot's behavior. When robot intent was transparent, participants reduced the number of demonstrations. This led to a decrease in participant workload and an increase in performance satisfaction. Feedback on their demonstrations further decreased both the number and distance of demonstrations, improving demonstration quality and further reducing participant workload, while also boosting confidence in the robot's capabilities. The results provide strong evidence that nonexperts can successfully train agents in new behaviors after deployment, and that greater transparency of the agent improves outcomes for both the human user and the robot.					
15. SUBJECT TERMS artificial intelligence/machine learning, AI/ML, intervention learning, agent transparency, trust, workload, Godspeed, human-autonomy teaming, Humans in Complex Systems, Military Information Sciences					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT		18. NUMBER OF PAGES
a. REPORT UNCLASSIFIED	b. ABSTRACT UNCLASSIFIED	c. THIS PAGE UNCLASSIFIED	UU		87
19a. NAME OF RESPONSIBLE PERSON Julia L. Wright				19b. PHONE NUMBER (Include area code) (616) 283-3829	

STANDARD FORM 298 (REV. 5/2020)

Prescribed by ANSI Std. Z39.18

Contents

List of Figures	vi
List of Tables	viii
Acknowledgments	x
1. Introduction	1
2. Background	2
2.1 Autonomous Agents	2
2.2 Agent Transparency	3
2.3 Feedback	4
2.4 Workload	4
2.5 Trust in Automated Systems	5
2.6 Human Perception of Agents	6
3. Study Methodology	6
3.1 Study Overview	6
3.1.1 Research Objectives	9
3.1.2 Hypotheses	9
3.2 Equipment and Facilities	10
3.3 Participants	11
3.4 Questionnaires	11
3.4.1 Demographics	11
3.4.2 Workload	11
3.4.3 Explicit Trust	12
3.4.4 Anthropomorphic Tendencies	12
3.4.5 Perception of Robot Learning	12
3.5 Experiment Design	12
3.5.1 Independent Variables	13
3.5.2 Dependent Measures	13

3.6	Procedure	14
4.	Results	14
4.1	Data Analysis Overview	14
4.2	Performance Measure Results	15
4.2.1	Increased Transparency Impacts Human Demonstrations	15
4.2.2	Confidence in Robot Increases over the Course of the Mission	17
4.2.3	Demonstration Performance Impacts Autonomous Performance	19
4.2.4	Increased Transparency Reduces Perceived Complexity	19
4.2.5	Transferability of Learned Models to New Missions	21
4.2.6	Performance Measures Regression Analyses	24
4.3	Subjective Measures	26
4.3.1	NASA-TLX	26
4.3.2	Trust in the Robot	28
4.3.3	Anthropomorphic Tendencies	29
4.3.4	Perception of Robot Learning	30
5.	Discussion	31
5.1	Synopsis	31
5.2	Limitations	33
5.3	Future Work	34
6.	Conclusion	34
7.	References	35
	Appendix A. Recruitment Materials	39
	Appendix B. Demographics Questionnaire	41
	Appendix C. NASA-TLX	43
	Appendix D. Trust in Automation	45

Appendix E. Godspeed Questionnaire Series	47
Appendix F. Robot Learning Questionnaire	49
Appendix G. Analysis Results Tables	51
List of Symbols, Abbreviations, and Acronyms	74
Distribution List	75

List of Figures

Figure 1.	Interface for transparency condition: transparency: baseline (T1). Left: learning condition (LC)1, T1. The left panel (top) shows the video stream from the robot’s onboard camera. The right panel (top) shows the perception information for the video stream, where each color represents a different terrain or object class. This perception information is also projected into the LiDAR-generated map view to illustrate the terrain with respect to the robot’s current location. Blue spheres represent the waypoints to where the robot will traverse. The top right shows a status message indicating the status of the robot. Right: LC2, T1. All visualizations remain the same, but the dangerous regions are also displayed to the participant. Dangerous regions are indicated by red spheres.....	7
Figure 2.	Interfaces for transparency condition: T2, transparent intent (left); T3, transparent impact (right). Left: LC1, T2. All visualizations from the T1 condition remain. The black trajectory in front of the robot indicates where the robot intends to traverse autonomously with its current behavior model. Right: LC1, T3. All visualizations from the T2 LC remain. In addition, information that indicates the trajectory driven during the demonstration is shown as a cyan line. Moreover, the top right panel shows the duration of the demonstration (in seconds) and the change from the robot’s originally planned path, indicating how different the demonstration is from the intended trajectory.	9
Figure 3.	Average number of demonstrations provided by participants across the three transparency levels. Bars denote SE.....	16
Figure 4.	Traversal behavior error, measured as the MHD compared to a set of ground-truth trajectories, for both demonstration and autonomous trajectories across the three transparency levels. Bars denote SE....	16
Figure 5.	Finer breakdown of the average number of demonstrations provided throughout the life of the mission for each of the three transparency levels. Bars denote SE.....	17
Figure 6.	Fine-grained breakdown of the autonomous traversal time for a robot across the life of the mission for each of the three transparency levels. Bars denote SE.....	18
Figure 7.	Fine-grained breakdown of the average demonstration distance (in meters) throughout the duration of the mission. Bars denote SE.....	18
Figure 8.	A comparison of the demonstration error and associated autonomous traversal error for the top performing participants compared to the bottom performing participants. Top performance is defined as participants with lowest demonstration trajectory error as computed by the MHD with respect to the ground-truth trajectories. Bars denote SE.....	19

Figure 9.	A fine-grained comparison of the average number of demonstrations provided across the life of the mission for the LC1 and LC2 participants. Although LC2 is more complex, fewer demonstrations were provided by the participants in the T2 and T3 cases. Bars denote SE.....	20
Figure 10.	Fine-grained breakdown of the demonstrations provided that were specific to learning to avoid dangerous regions in the LC2 cases. As the mission progresses, fewer of these demonstrations are provided (as seen across all transparency cases), indicating that this portion of the behavior was learned well in the earlier portions of the mission. Bars denote SE.	21
Figure 11.	Qualitative trajectory results for several models deployed in a novel LC1 trial compared to those tested in the study. Deployed models are those produced by top performing participants in each of the three transparency levels and the initial behavior model deployed at the start of each trial. Notice that the model trajectories better resemble the ground truth trajectories as more transparency is introduced.....	22
Figure 12.	Qualitative trajectory results for several models deployed in a novel LC2 trial compared to those tested in the study. Deployed models are those produced by top performing participants in each of the three transparency levels and the initial behavior model deployed at the start of each trial. Notice that the model trajectories better resemble the ground truth trajectories as more transparency is introduced.....	23
Figure 13.	Quantitative measures of the traversal error for the novel mission introduced. Bars denote SE.	24
Figure 14.	NASA-TLX scores, overall and by subfactor. Bars denote SE.....	26
Figure 15.	LC1 NASA-TLX scores. Bars denote SE.	27
Figure 16.	LC2 NASA-TLX scores. Bars denote SE.	28
Figure 17.	Jian TIA scores. Bars denote SE.	28
Figure 18.	Overall GQS scores, by trial. Bars denote SE.....	29
Figure 19.	Perception of RLQ scores across transparency conditions (T1, T2, and T3). Bars denote SE.....	30
Figure 20.	Perception of RLQ scores for LC1 and LC2 across transparency conditions (T1, T2, and T3). Bars denote SE.....	31

List of Tables

Table G-1.	Standard multiple regression of performance outcome variables on NASA-TLX scores.....	52
Table G-2.	Standard multiple regression of performance outcome variables on TIA survey scores.....	54
Table G-3.	Standard multiple regression of performance outcome variables on Godspeed series survey scores.	55
Table G-4.	Standard multiple regression of performance outcome variables on robot learning questionnaire scores.	56
Table G-5.	Performance measure correlations with confidence intervals.	58
Table G-6.	Overall descriptive statistics for NASA-TLX questionnaire results.	59
Table G-7.	Repeated measures ANOVA results for overall NASA-TLX scores between trials.....	59
Table G-8.	Post hoc paired t-test results for overall NASA-TLX scores.	60
Table G-9.	NASA-TLX scores descriptive statistics by learning condition (LC).	60
Table G-10.	NASA-TLX paired t-test results, within learning condition (LC). ..	61
Table G-11.	NASA-TLX factor correlations with performance measures.....	61
Table G-12.	Overall Jian TIA descriptive statistics.	63
Table G-13.	Repeated measures ANOVA results for Jian TIA questionnaire scores.	63
Table G-14.	Post hoc paired t-test results for Jian TIA overall scores.	64
Table G-15.	Jian Trust in Automation descriptive statistics, by learning condition (LC).	64
Table G-16.	Jian TIA paired t-test results within learning condition (LC).	64
Table G-17.	Jian TIA measures correlations with performance measures.....	65
Table G-18.	Overall Godspeed Questionnaire Series descriptive statistics.	66
Table G-19.	Repeated measures ANOVA results for Godspeed Questionnaire Series scores	66
Table G-20.	Post hoc paired t-test results for Godspeed Questionnaire Series overall scores.....	66
Table G-21.	GQS descriptive statistics, by learning condition (LC).....	67
Table G-22.	Post hoc paired t-test results for GQS scores by learning condition (LC).	67
Table G-23.	GQS measure correlations with performance measures.	68

Table G-23. GQS measure correlations with performance measures (continued).	69
Table G-24. RLQ scores, descriptive statistics.....	69
Table G-25. Robot learning scores, repeated measure results.....	70
Table G-26. Robot learning scores, paired t-test results.....	70
Table G-27. Robot learning scores descriptive statistics by learning condition (LC).	70
Table G-28. Robot learning scores, paired t-test results by learning condition (LC).	71
Table G-29. RLQ factor correlations with performance measures.....	71

Acknowledgments

The authors wish to thank Joshua Foldes (University of Central Florida) for his contributions to this project.

1. Introduction

The ability of autonomous agents to conduct tasks of ever-increasing complexity is developing at an exponential rate with the use of artificial intelligence (AI) and machine learning (ML). The advancements of deep learning that propelled visual perception performance (Alom et al. 2018) and recent successes of large language models (Radford 2018) showcase the reliance that AI/ML algorithms have on large, in-distribution datasets to learn effectively. Although making use of large datasets for learning is an ideal scenario, not all applications have this luxury, including those related to off-road autonomous ground navigation. For off-road navigation tasks, datasets are relatively sparse (Jiang et al. 2021) (compared to the vast amount of imagery and text available on the Internet), capture only a small portion of the globe, and are primarily focused on capturing perception-relevant features of a scene with less emphasis on providing explicitly meaningful control output to learn specific behavior (e.g., traversal tactics and protocols).

Leveraging existing data allows for initial model training for an autonomous agent; however, when this is deployed to highly dynamic or novel environments, agent autonomy models will often fall short of meeting ever-changing needs and newly arising situations if they are not equipped with the ability to update regularly. One solution to this problem is the development of learning-capable agents, that is, agents that can learn and adapt their behavior over time based on past experiences, updated world models, and instruction or feedback. By continuously adapting to new information, learning-capable agents ensure their reasoning and actions suit the evolved needs of the mission.

One approach to universal operational capability is to augment autonomous systems training with human demonstration at the beginning of an operation in novel surroundings (Wigness et al. 2018). Although this approach provides opportunities to collect domain-specific data for the expected mission execution, there is no guarantee there will be sufficient time or access to all relevant portions of the environment to collect the needed data. Thus, it may still be difficult (if not impossible) to deploy the autonomous agent with all possible behavioral knowledge that will be required for dynamic mission execution.

In these cases, it becomes critical that human interventions can be provided and leveraged during operation when faced with changing environments or deteriorated behavior (Goecks et al. 2019) to ensure continued autonomous agent execution. This technical report discusses the research exploring the scenario in which an autonomous agent with a learned navigational model must change its behavior after deployment with the assistance of a human operator. We examine the efficiency

and efficacy of the updated agent model, as well as the effect training the agent has on the human operator, and we explore information exchanges that could potentially influence the teaming.

2. Background

We briefly define and discuss relevant concepts in the human-agent teaming literature that are relevant to this research. We outline related work in these areas and the existing scientific gaps that this research addresses.

2.1 Autonomous Agents

Autonomous systems can make decisions, select courses of action, direct their activities to achieve a goal, and use specified criteria to evaluate whether they are successful (Wooldridge and Jennings 1995; David and Nielsen 2016; Lyons et al. 2018; Wynne and Lyons 2018; O’Neill et al. 2020). When these systems are also capable of learning, they are not static entities; they use their experiences to update the models that guide their actions and decision making. Learning frameworks that produce the best performance require a supervisory signal to build models, and in human-agent teaming domains this learning signal can be provided through human feedback, such as providing reinforcement or advice (Breazeal and Thomaz 2008) or behavior demonstrations (Wigness et al. 2018).

In the robotics research community, learning-based approaches for autonomous navigation represent a large body of literature (Xiao et al. 2022a) that showcase the utility of learning from demonstration (Wigness et al. 2018), reinforcement learning (RL) (Luong and Pham 2021), and end-to-end deep learning (Kahn et al. 2021a) to accomplish a variety of navigation tasks. In most cases, these approaches learn offline, using batches of training data (e.g., demonstrations, RL epochs, or labeled samples) to identify a reward function or policy before deployment.

The performance of these ML algorithms is often highly dependent on the amount of the training data provided for learning. Too few data could result in overfitting to specific data instances because the data is unable to adequately represent the underlying feature space. However, large sets of redundant data also suffer from underrepresentation of the feature space for learning. At deployment, the learned models and policies remain static and are used to determine the next action the robot should take given its current environment state. In cases in which deployment conditions are dynamic because of changes in mission requirements or platform capabilities, existing approaches fail to provide the adaptation needed to maintain accurate, reliable, and safe navigation. Thus, the ability to dynamically update

behavior models provides autonomous agents with some resiliency in the field when faced with unexpected conditions.

Xiao et al. (2022b) showed that with Adaptive Planner Parameter Learning, multiple sets of parameters could be learned for different contexts and adaptively loaded at runtime based on a context classifier. However, this set of parameters remains static and only provides the necessary information for the predefined contexts. Kahn et al. (2021b) used Learning to Navigate from Disengagements to collect data from a human when they believed that an autonomous vehicle needed to be disengaged, that is, they intervened autonomous operation to ensure safety of the vehicle and surrounding objects. This information was used to retrain a model that could predict when the trajectories would likely result in disengagements, resulting in higher quality navigation. This fine-tuning with disengagement information was also collected in a batch and performed in an off-line fashion. Thus, there is a gap in the literature that makes use of online, post-deployment training, where an agent’s behavior model is dynamically adapted and improved to reflect changing goals, objectives, or restrictions, while conducting tasks in the field without incurring downtime to accommodate off-line learning. Our study explores this objective using real-time human intervention during autonomous operation to provide demonstration feedback for behavior model updating. We modify the off-line inverse algorithm previously used in robotic field experimentation (Wigness et al. 2018) to run and retrain behavior models in real time after each human intervention.

2.2 Agent Transparency

Transparency of autonomous systems, the ability of an autonomous system to provide information about its internal workings and decision-making processes to users, operators, or other stakeholders, is a topic that directly impacts human–robot interactions. Agent transparency, as a tool to enhance users’ understanding of the agent’s abilities, intent, reasoning, performance, and plans within the context of its goals (Lyons 2013; Chen et al. 2014), significantly influences trust, accountability, and safety. By enabling users to comprehend the agent’s decision-making processes and how it arrives at its decisions (Chen et al. 2014), transparency can help users anticipate the agent’s behavior and make informed decisions about when and how to collaborate with the agent (Barber et al. 2015). Transparency has also been shown to improve team performance (Selkowitz et al. 2017; Wright et al. 2024), situational awareness (Chen et al. 2014; Selkowitz et al. 2016), and supports trust calibration (Hoff and Bashir 2015; Schaeffer et al. 2017). In this study, we examine how adjusting the level of transparency by adding information about the agent’s

intent and the participant's demonstrations influences the human's demonstrations and perception of the robot.

2.3 Feedback

Feedback is information about one's performance or understanding of a concept, which is usually provided by an external source, but it can also be self-provided via experience (Hattie and Timperley 2007). Feedback can refer to a task, a process, self-regulation (self-directing or directing), or self-level (personal evaluations and affect). The most impactful feedback relates to a task or a process; on the other hand, self-feedback is not only the least impactful, but it can dilute the effectiveness of other forms of feedback when used in conjunction with them. There are several variations of feedback: "feed-up" compares actual status to a target or goal; "feedback" compares an actual status to a previous status; and "feed-forward" occurs when the objective is explained in terms of the current status, often resulting in a change in strategy or processes (Hattie and Timperley 2007). Feedback is most effective when the information content is detailed, the desired goals are specific, and task complexity is low (Hattie and Timperley 2007; Wisniewski et al. 2020). In this study, the highest level of transparency, impact transparency, yields information about the participant's demonstrations relative to the robot's planned path (i.e., feed-up information). How this feedback influences the number and quality of participant demonstrations is a key objective of this study.

2.4 Workload

The concept of workload, as per attention resource theory (Wickens 2002), refers to the relationship between the resources required by a task and the resources available to the individual performing the task. This is a subjective construct, as two individuals performing the same task or using the same systems can produce identical outputs, but one person may have more attention to spare for concurrent tasks than the other. An individual's cognitive workload assessment can be based on performance, subjective reports, or physiological measures. The aim of workload assessment is to ensure that the benefits to the individual outweigh the costs incurred from the working environment.

In the human–autonomy teaming context, cognitive workload is heavily influenced by the levels of automaticity and transparency. In a recent study, a participant teamed with a learning-capable target classification agent to identify potential threats (Wright et al. 2024). The agent shared no information as to its reasoning in the baseline condition. In each successive condition, transparency increased to indicate the agent's reasoning and again to indicate how the agent interpreted

information supplied by the participant. In the highest transparency level, the participant was required to indicate the specific logic factor the agent was missing or could not determine. In the first three conditions, it was demonstrated that increased agent transparency reduced workload, specifically mental demand, temporal demand, effort and frustration, and participants' performance satisfaction. However, when the participant was responsible for directly instructing the agent when it indicated it had insufficient information to complete the classification, workload ratings were as much or higher than in the baseline condition, and performance satisfaction was at its lowest (Wright et al. 2024). This work demonstrated that operator workload can still decrease as shared information increases; however, increasing the operator's tasking will counteract the benefits of increased transparency (Parasuraman et al. 2000). In this study, we will examine how increasing the information available to the participant supports the operators' workload when their primary task is to intervene and instruct the agent.

2.5 Trust in Automated Systems

An operator's trust in an agent is a dynamic state, continuously being reassessed and calibrated during their interaction (de Visser et al. 2019). This trust is influenced by many factors, both internal, such as operator's attitudes, propensity to trust, operator experience, and external, such as agent reliability, performance, and transparency (Merritt et al. 2013; Chen et al. 2014, 2018; Drnec and Metcalfe 2016). Calibrated trust is essential for effective team performance in human-agent teams (Jian et al. 2000; Hoff and Bashir 2015; Schaefer et al. 2016) because it enables appropriate use of the agent and maintenance of a team's effectiveness (Parasuraman and Riley 1997).

In agents with static behavior models, transparency of the agent's actions, intents, and processes supports an operator's appropriate trust in the agent (Lee and See 2004; Chen et al. 2014, 2018). However, there is limited research into how transparency supports trust when the agent can learn and update its behavior during deployment. In a study in which a participant teamed with a simulated learning agent on a target classification task, the transparency information supplied to the participant focused on what information the agent needed to complete a classification. When the participant was required to intervene and give the agent specific information it was missing, their reported trust in the agent significantly worsened (Wright et al. 2024), suggesting that the participant's increased awareness of the agent's shortcomings had the effect of reducing their trust in the agent. In this study, the participant's role is to teach an agent (robot) a new navigation behavior. The transparency information focuses on the agent's actions and intent and the impactfulness of the participant's demonstration of the desired

behavior. We expect that increased transparency will support appropriate trust in the agent.

2.6 Human Perception of Agents

How the human operator perceives the robot is of interest in this study. Users differentiate between autonomy and people in performance expectations (Merritt et al. 2013) and assigning intent to the robot (Lee and See 2004). However, when users perceive a robot as more humanlike, they are more willing to cede responsibility for the task outcomes to the robot (Hinds et al. 2004). In a prior study with learning-capable agents, when the participants were tasked with supplying an agent with the information it was missing to make a threat categorization, their perceptions of the agent were more negative. They viewed the agent as having less animacy, likability, and intelligence (Wright et al. 2024). In this study, the participants' primary task is to teach an autonomously navigating robot a new behavior; as such, more negative perceptions of the robot could lead to participants engaging in over monitoring behavior (Hinds et al. 2004).

3. Study Methodology

3.1 Study Overview

In this experimental study, participants played the role of a human teammate overseeing the autonomous actions of an unmanned ground robot driving through a semi-structured village environment. The human teammate role was designed such that they could only remotely observe the robot using raw sensor feeds and maps generated by the robot online. Figure 1 shows the baseline user interface. The top left includes a video stream from the robot's onboard camera sensor. Directly to the right of the raw video stream is the robot's perceptual understanding of this data where semantic classes (e.g., building, road, or tree) are color coded. The remaining portion of the interface is composed of an overhead map of the operational area that defines obstacles using the Light Detection and Ranging (LiDAR) sensor, the color-coded terrain semantics (pink for road and green for grass in Figure 1), mission waypoints (blue circles in Figure 1), and dangerous regions, when applicable (red circles in Figure 1).

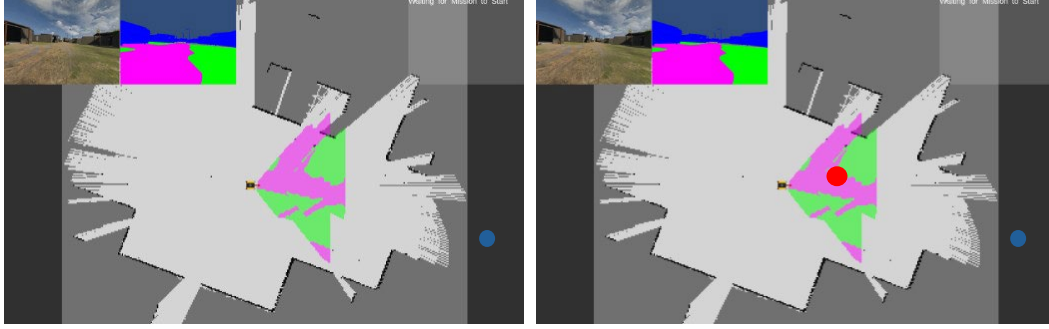


Figure 1. Interface for transparency condition: transparency: baseline (T1). Left: learning condition (LC)1, T1. The left panel (top) shows the video stream from the robot’s onboard camera. The right panel (top) shows the perception information for the video stream, where each color represents a different terrain or object class. This perception information is also projected into the LiDAR-generated map view to illustrate the terrain with respect to the robot’s current location. Blue spheres represent the waypoints to where the robot will traverse. The top right shows a status message indicating the status of the robot. Right: LC2, T1. All visualizations remain the same, but the dangerous regions are also displayed to the participant. Dangerous regions are indicated by red spheres.

The robot is tasked to autonomously traverse a four-waypoint mission, where the robot will plan paths to each waypoint in the specified visitation order. The robot is equipped with a behavior traversal model that uses the LiDAR and semantic terrain information to plan paths that represent a high-level behavior (e.g., drive on the road). For this study, the initial behavioral model that is deployed at the start of each trial is one that exhibits a slight preference for traversing over grass terrain. However, the details of the initial deployed model are not provided to the participants. Furthermore, the traversal behavior that participants are instructed to train the robot to execute differs substantially from the base policy of the initially deployed model.

Each participant is given details about the specific traversal behavior the robot should be executing. The participant’s task is to remotely observe the robot’s actions, and if their robotic teammate is performing the specified traversal behavior in an unsatisfactory manner, they are to intervene in the robot’s autonomous operation and demonstrate the desired traversal behavior by teleoperating the robot using an Xbox controller. When to intervene and the length of the demonstration is determined entirely by the participant. Each provided demonstration is then immediately used by the robot to update its traversal behavior model to better mimic the behavior depicted in the trajectory provided by the human. The model update process involves the robot pausing all autonomous or user demonstration operations and using up to 60 s to optimize the underlying model parameters given the new data. The trial then continues with the robot leveraging the newly updated behavior model, while the human teammate’s task remains unchanged.

The study explores two traversal behaviors for the robot, which we refer to as LCs:

- LC1: Planned traversal trajectories should stay on the road while maintaining proximity to the edge of the road.
- LC2: Planned traversal trajectories should stay on the road while maintaining proximity to the edge of a road unless a dangerous region prohibits this, in which case the trajectory should deviate from this behavior to avoid the dangerous region before returning to the edge of the road.

The LC2 behavior was designed to negotiate between two distinctive sub-behaviors, i.e., following the edge of the road and avoiding dangerous regions. This negotiation between two sub-behaviors results in a more complex learning objective, which allows us to examine how increased complexity impacts the research objectives.

In addition to the two LCs, there are three transparency (T) conditions explored in the study:

- T1: Baseline
- T2: Transparent intent; the robot shares its intended path to the following waypoint
- T3: Transparent impact; the robot assesses how different the demonstration is from the current model and conveys that information to the participant

The displays for the three transparency conditions (T1–T3) are shown in Figures 1 and 2. The utility measure in T3 is computed using the modified Hausdorff distance (MHD), which computes the distance between poses in the demonstrated trajectory provided by the participant compared to the robot’s intended autonomously planned path. Larger values indicate a greater difference between the two paths, suggesting the demonstration will be more impactful when used to update the current robot behavior model.

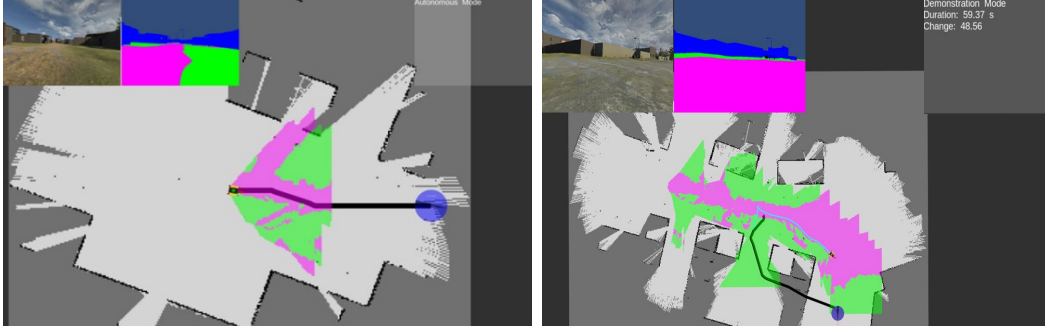


Figure 2. Interfaces for transparency condition: T2, transparent intent (left); T3, transparent impact (right). Left: LC1, T2. All visualizations from the T1 condition remain. The black trajectory in front of the robot indicates where the robot intends to traverse autonomously with its current behavior model. Right: LC1, T3. All visualizations from the T2 LC remain. In addition, information that indicates the trajectory driven during the demonstration is shown as a cyan line. Moreover, the top right panel shows the duration of the demonstration (in seconds) and the change from the robot’s originally planned path, indicating how different the demonstration is from the intended trajectory.

Participants were assigned to one LC and completed three trials, one in each of the transparency conditions. For each trial, one of three random mission routes were selected for operation that defined the four waypoints the robot must traverse. Questionnaires (discussed in greater detail in Section 3.4) were administered to the participants both during the trial (at waypoints) and at the completion of the trial. During the trial, the robot would pause autonomous operation at each waypoint and the participant was given a short survey to answer questions about the robot’s behavior. At the final waypoint of the trial, the participant completed several questionnaires on their workload, trust, and impressions of the robot.

3.1.1 Research Objectives

The goals of this research were as follows:

- 1) To investigate the effectiveness of the ML algorithm when provided with human demonstrations to facilitate online learning.
- 2) To investigate how transparency of the demonstration impact affects the creation of demonstrations that are sufficient to adequately update the existing algorithm.
- 3) To investigate the effect of transparency of a robot path planning system on participant workload, trust, and attributions of the robot.

3.1.2 Hypotheses

Hypotheses for the study were generated based on a review of the relevant literature. In the following section, the rationale for each hypothesis will be discussed, and then the resultant hypothesis will be stated. The goal of this research

is to improve robot learning in the field through intervention and demonstration. As such, we predict that when the human teammate performs “better” demonstrations (e.g., longer duration, clearer intent, significantly different from the robot’s intended planned path), then robot learning will improve (hypothesis 1). Therefore, the performance of the robot to mimic the desired traversal behavior will improve as the quality of human demonstration increases.

We predict that the human demonstration will be dependent on the complexity of the behavior they want the robot to learn (hypothesis 2). Since LC2 is more complex, with multiple behavior objectives to learn compared to just one in LC1, the human will intuitively adjust their demonstrations appropriately. As a result, LC2 will have more frequent and longer human demonstrations than LC1.

Communication between the robot and the human teammate is expected to be central to achieving the goal of robot learning. When the human understands what the robot is doing and why, as well as how their demonstration is interpreted by the robot, the result should be improved teaming and performance outcomes. In this study, team communications are supported by the robot interface and display (i.e., transparency). As such, we predict that as transparency in the interface increases (i.e., T1 to T2, T2 to T3), the following will occur (hypothesis 3):

- a) Improved effectiveness of human demonstration (i.e., higher quality)
- b) Reduced cognitive workload
- c) Increased human trust in the robot
- d) Increased perception of the robot’s animacy, intelligence, and likeability
- e) Perception of greater robot learning

3.2 Equipment and Facilities

The study was conducted in an indoor, climate-controlled office space. The U.S. Army Combat Capabilities Development Command Army Research Laboratory’s Ground Autonomy Stack defined the software components of this experiment that enabled autonomous agent learning and behavior execution. We used a DEVCOM Army Research Laboratory Unity-based simulation environment that represented a semi-structured village with several buildings, trees, and an unpaved road network. All processing and visualization of the user interface was delivered via a commercial computer system, using a 27-inch computer monitor and a standard Xbox controller. Questionnaires were administered via a standard commercial laptop.

3.3 Participants

Forty-five participants were recruited via the university's participant recruitment system (Appendix A), which is open to the public. Participants were compensated \$15/h for their time in the study, rounded up to the nearest hour, and paid in gift cards. Participants could withdraw from the study at any time. Four of the 45 participants were removed from the study; three due to simulator malfunctions and one person withdrew. Two of the simulator errors occurred at the end of the final trial. As such, performance data (i.e., data captured from the participant's interactions with the simulation software) from 43 participants were analyzed. Subjective data from 41 participants were analyzed, because experiencing the simulator malfunction was expected to influence the participant's responses.

3.4 Questionnaires

All questionnaires were administered on a laptop computer using the Psychology Experiment Building Language (PEBL) software version 2.1 (Mueller and Piper 2014; Mueller 2014).

3.4.1 Demographics

A demographics questionnaire was administered before the training session began (Appendix B). Information on the participants' age, gender, military experience, and AI/ML experience was collected. Thirty-two participants were 18–25 years old, six participants were 26–35 years old, and three participants were 36–45 years old. Twenty-four participants reported their gender as female and 17 participants reported their gender as male; none were undisclosed or preferred not to answer. Five participants had prior military experience, none of whom operated robots or drones as part of their duties. Thirty-five participants reported having no experience with AI/ML or being somewhat familiar with terms and concepts only, five participants reported being somewhat experienced, and one participant reported being an expert in an area other than robotics.

3.4.2 Workload

Participant's subjective workload was assessed using the computerized version of the NASA-Task Load Index (TLX) questionnaire (Hart and Staveland 1988). The NASA-TLX is a self-report questionnaire of perceived demands in six areas: mental demand, physical demand, temporal demand, effort (mental and physical), frustration, and performance. Participants evaluated their perceived workload on these factors using Likert-type scales ranging from 0 (Low, Good) to 100 (High,

Poor) at the end of each trial. Overall workload was assessed using an average of all raw factor scores (Appendix C).

3.4.3 Explicit Trust

At the end of each trial, participants assessed their trust in the robot using a modified version of the Trust in Automation (TIA) scale (Jian et al. 2000). This scale comprises 12 items, such as “The robot is deceptive,” which the participant assesses using a seven-point Likert scale (1 = not at all, 7 = extremely), wherein higher scores indicate more trust in the agent (Appendix D). The scores were calculated as separate factors for “distrust” (questions 1–5) and “trust” (questions 6–12) rather than a single aggregate score (Wright et al. 2024).

3.4.4 Anthropomorphic Tendencies

The Godspeed Questionnaire Series ([GQS] Bartneck et al. 2009) comprises 24 questions in five scales that evaluate user opinions of social aspects of a robot during a human–robot interaction task: anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety. The GQS items (Appendix E) are rated using a five-point Likert scale with endpoints that are semantic differentials (e.g., awful/nice). Internal reliability consistency ranges from 0.702 (animacy) to 0.929 (anthropomorphism). Participants completed a Godspeed measure after each trial. The questions were presented to the participants in a random order.

3.4.5 Perception of Robot Learning

The Robot Learning Questionnaire ([RLQ] Appendix F) is a novel questionnaire developed by the researchers of this study. This questionnaire assesses the participant’s perception of the robot’s learning and future performance. This survey was administered each time the robot arrived at a mission waypoint (i.e., four surveys per trial), and the participant rated how well the robot had learned the desired behavior using a four-point Likert scale (1 = not at all, 4 = completely trained).

3.5 Experiment Design

The study was a 2 (LC1, LC2) $\times$ 3 (T1, T2, T3) mixed-factor design. Within-subjects evaluations compared differences in performance, workload, trust, and attributions regarding the robot across levels of transparency of robot intent and demonstration effectiveness, both across and within LCs. The participant’s LC assignment was randomized and counterbalanced. The transparency conditions were presented in sequence; however, the specific test paths were randomized across conditions. The within-subjects factor (transparency) was evaluated using

repeated measures analysis of variance (ANOVA) with planned post hoc contrasts. Means, standard deviations (SDs), and effect sizes are reported for all measures and tests.

3.5.1 Independent Variables

The independent variables were LC (LC1, LC2) and transparency level (T1, T2, T3). The two LCs explored how task complexity impacts the research objectives. The three transparency conditions examined how varying the participant's information relating to the robot's behavior and the impact of their own demonstration affected performance and perceptions. Detailed descriptions of the independent variables are in Section 3.1.

3.5.2 Dependent Measures

Dependent variables are found within the performance data captured from the participants' interactions with the simulation software and the questionnaires provided both during and after each trial. From the performance data, we assess the following variables associated with both the demonstrations provided and resulting autonomous behavior:

- NumDemos: number of demonstrations provided
- Distance: length of demonstrations (measured in meters)
- Duration: duration of demonstrations (measured in seconds)
- Quality-MHD: demonstration trajectory error with respect to ground-truth trajectories that depict the desired behavior outlined in the task. Measured as the MHD compared to the ground truth.
- AutoQuality-MHD: autonomous trajectory error with respect to ground-truth trajectories that depict the desired behavior outlined in the task. Measured as the MHD compared to the ground truth.
- AutoDistance: the percentage of total distance traveled that the robot operated autonomously.

The ground-truth trajectories used to compute some of the measures were collected by two different researchers. Multiple ground-truth trajectories were collected since there could be multiple ways for the robot to execute the intended behavior (e.g., the robot could avoid a dangerous region in LC2 by driving around it in either a clockwise or counterclockwise direction). Each demonstration or autonomous trajectory measure is computed by comparing to the most similar ground-truth trajectory.

From the subjective data, we assess variables from each of the four questionnaires completed by the participant during and after each trial: the NASA-TLX, modified

Jian et al. (2000) TIA survey, GQS, and RLQ. Details of these surveys are in Section 3.4.

3.6 Procedure

Participants were recruited via the university's Sona research recruitment system. When the participants reported to the designated time and place for the study, they reviewed and signed the informed consent form and then completed the demographics questionnaire.

The participant was shown to the experimental simulation station and learned how to use the display content and Xbox controller. Then, they completed self-paced training (via PowerPoint slides) and a practice run, which consisted of having the participants drive the robot in the simulation environment using the controller and familiarizing them with the initial interface display. Participants were allowed to ask questions during training. The participants were quizzed at the end of their initial training to ensure they could operate the robot in the simulation, understood the displayed information, and understood the robot learning objective. Between each trial, participants had a brief training via PowerPoint slide on the added transparency features of the display. Again, they were allowed to ask questions about the new visual content. They then completed a short quiz demonstrating their understanding of the new material.

Participants completed three trials in their assigned LC, one in each transparency condition. Participants received short breaks (maximum 5 min) between trials. At each waypoint during and after each trial, participants answered questions about the robot and completed questionnaires regarding their trust, workload, and perspectives of the robot. After completing all trials, participants were debriefed, and the experimenter answered any questions about the study. They were then thanked and dismissed.

4. Results

4.1 Data Analysis Overview

Subjective data were analyzed using repeated measures ANOVAs ($\alpha = .05$) and planned comparisons (T1 to T2, T2 to T3). When appropriate, the Greenhouse–Geisser correction for sphericity violation is reported. Post hoc planned comparisons were conducted using paired t-tests. Partial eta-squared (η^2_p) and Cohen's d effect sizes are reported when applicable. Descriptive statistics and t-test result tables are in Appendix G.

The dependent measures of the performance data were computed for each of the transparency trials. Comparison of these measures across trials was done by analyzing the mean and standard error (SE) across the participant data. In some cases, a finer-grained analysis of the performance data is computed to evaluate how performance changes within a trial. That is, the dependent measures are computed between each waypoint in the mission for each trial.

4.2 Performance Measure Results

Analysis of the performance measures yields four high-level takeaways from the study related to 1) the impact of increased transparency, 2) the participant’s confidence in the robot’s ability to learn, 3) the impact of demonstration quality on autonomous operation, and 4) the impact of transparency on perceived complexity of a problem. We detail the performance results leading to these takeaways and discuss transferability of the learned models from the study. Subjective data that supports these findings are in Section 4.3.

4.2.1 Increased Transparency Impacts Human Demonstrations

We begin by analyzing the performance data that directly relates to the demonstrations provided by the participants. By evaluating the average number of demonstrations that participants provided, we can determine if the transparency level had an impact on the number of demonstrations the participants believed they needed to provide to adequately train the robot to meet the specified traversal objective.

Figure 3 shows that, on average, fewer demonstrations were provided by participants as the transparency level increased. This indicates that the added visuals in T2, to better understand the robot’s traversal intent, and in T3, to illustrate the potential demonstration impact, allowed the humans to reduce their demonstration intervention. More specifically, the T2 visualization of robot intent (i.e., intended planned path) provides participants with information for a longer time horizon that likely makes it easier to determine whether the robot plan adheres to the specified LC behavior. However, in T1, participants only have visual information from the current time step, which likely leads to the participants intervening more often and potentially when unnecessary.

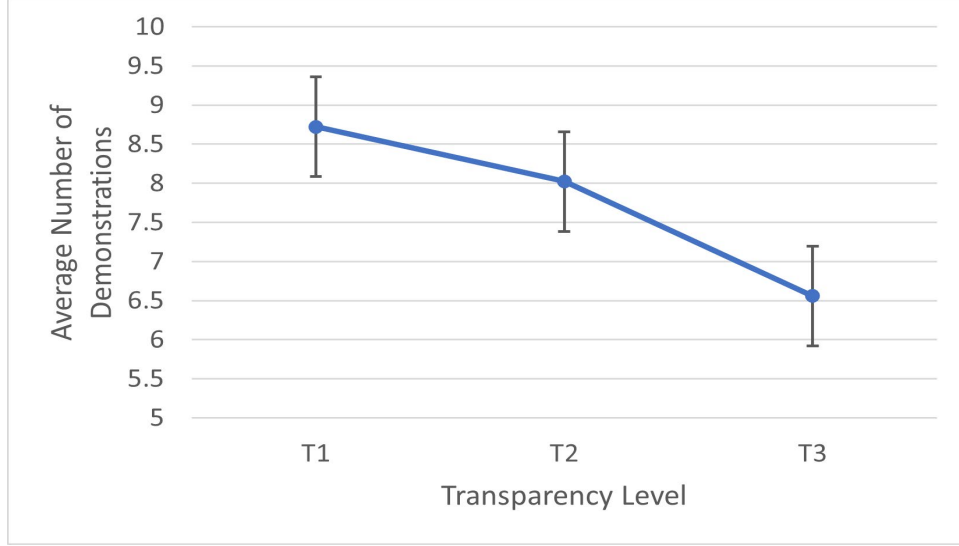


Figure 3. Average number of demonstrations provided by participants across the three transparency levels. Bars denote SE.

The additional visuals provided in T2 did not affect demonstration quality; however, increased transparency in T3 led to better quality demonstrations, supporting H3A. This quality improvement has an impact on learning and results in fewer data samples to adequately learn the intended traversal behavior. Figure 4 shows data to support this, illustrating that the trajectory error of the demonstrations decreases significantly for T3 ($F(2, 41) = 5.96$, $p = 0.005$, $\eta^2_p = 0.225$); however, the autonomous error stays relatively similar across all transparency levels. Thus, fewer high-quality demonstrations can provide the same learning benefits as more demonstrations with slightly less quality.

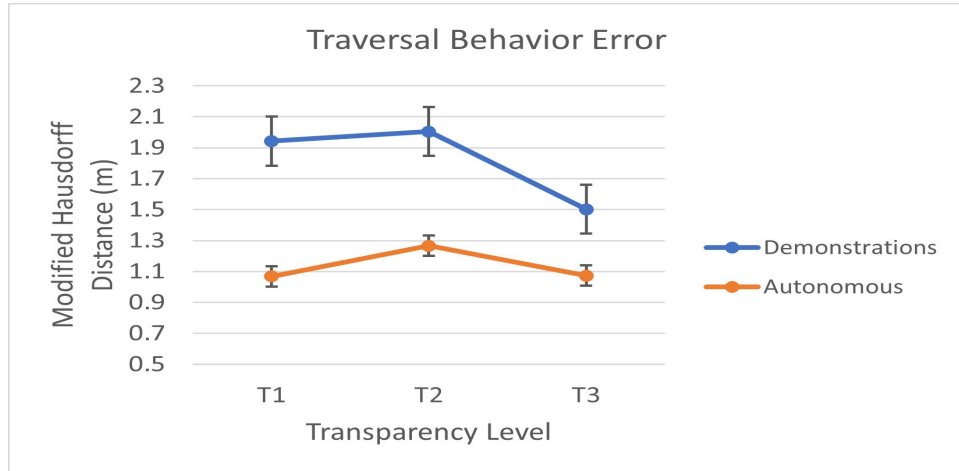


Figure 4. Traversal behavior error, measured as the MHD compared to a set of ground-truth trajectories, for both demonstration and autonomous trajectories across the three transparency levels. Bars denote SE.

4.2.2 Confidence in Robot Increases over the Course of the Mission

In each trial, the objective for participants is to evaluate the robot’s autonomous traversal behavior and intervene to provide demonstrations if the observed behavior does not adequately represent the LC behavior. Next, we look at the performance data to determine whether there are any data indicating the participant had confidence in the robot’s ability to learn.

We begin by evaluating the number of demonstrations provided at a finer granularity for each trial. Each mission is run in the simulation environment with four waypoints the robot must traverse in order, and Figure 5 breaks down the number of demonstrations that are provided between each of these waypoints. The results show that fewer demonstrations are provided when the robot traverses to the last two waypoints (labeled C and D) compared to those provided at the beginning of the mission to the first two waypoints (labeled A and B). Regardless of transparency level, the participants believed that the robot needed less guidance after making it through the first half of the mission, indicating the belief that the robot was learning.

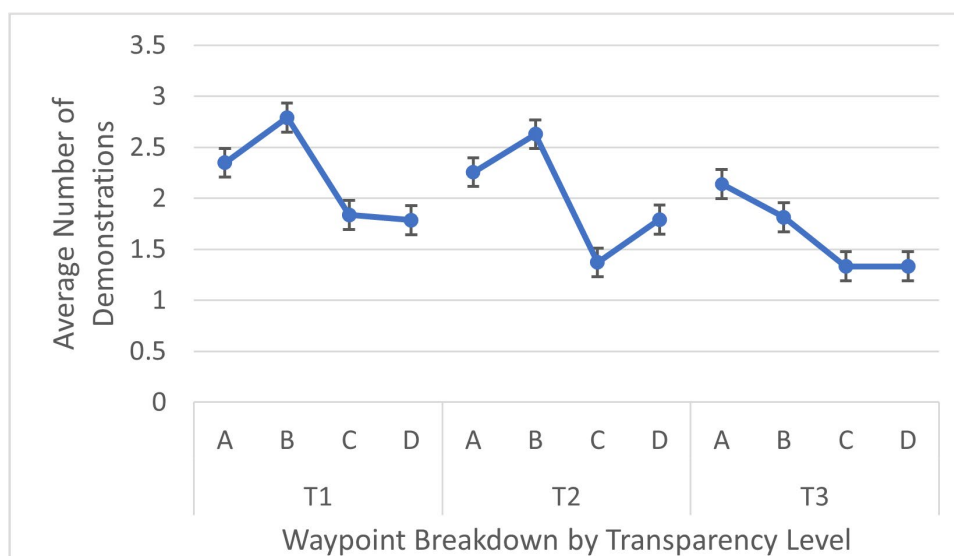


Figure 5. Finer breakdown of the average number of demonstrations provided throughout the life of the mission for each of the three transparency levels. Bars denote SE.

A decrease in the number of demonstrations alone does not provide adequate information to deduce that the participant had higher confidence in the robot later in the mission. For example, participants could have chosen to intervene immediately after the robot departed a waypoint and provided a demonstration the entire distance to the next waypoint. However, the performance data show that this is not the case. Figure 6 illustrates the percentage of the time the robot operated autonomously between each waypoint. Across each transparency level, the

participant allowed the robot to operate autonomously for longer periods of time as the mission progressed.

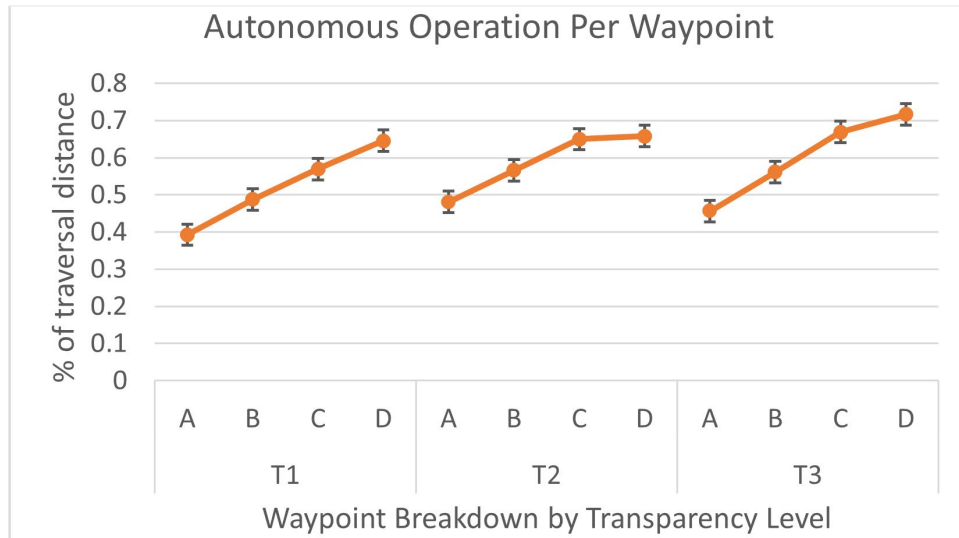


Figure 6. Fine-grained breakdown of the autonomous traversal time for a robot across the life of the mission for each of the three transparency levels. Bars denote SE.

This is partially because fewer numbers of demonstrations were provided; however, Figure 7 also shows that on average, the demonstrations became shorter as the mission progressed. This indicates that demonstration data collected near the end of the mission may have only been necessary to fine-tune an already high-performing behavior model.

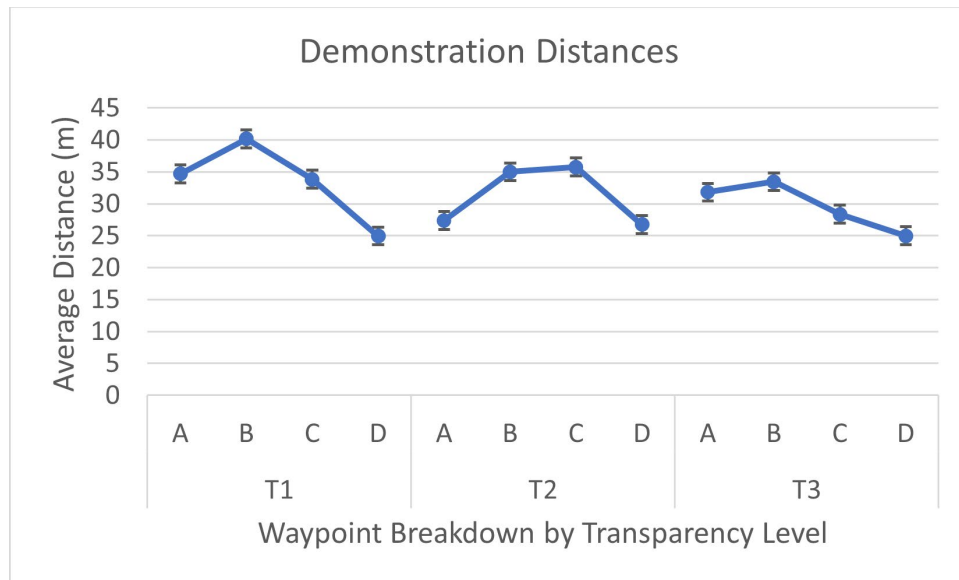


Figure 7. Fine-grained breakdown of the average demonstration distance (in meters) throughout the duration of the mission. Bars denote SE.

4.2.3 Demonstration Performance Impacts Autonomous Performance

Our first hypothesis (discussed in Section 3.1.2) stated that the robot’s autonomous performance would improve as the human demonstration performance improved. To evaluate this hypothesis with the performance data, we break down the participants into the top 50% of performers versus the bottom 50% of performers based on their demonstration error. Figure 8 shows that there is a performance gap in the autonomous behavior between these two groups of participants. The autonomous performance of the robot is better (i.e., lower error) when provided with demonstrations that are of higher quality and have lower error.

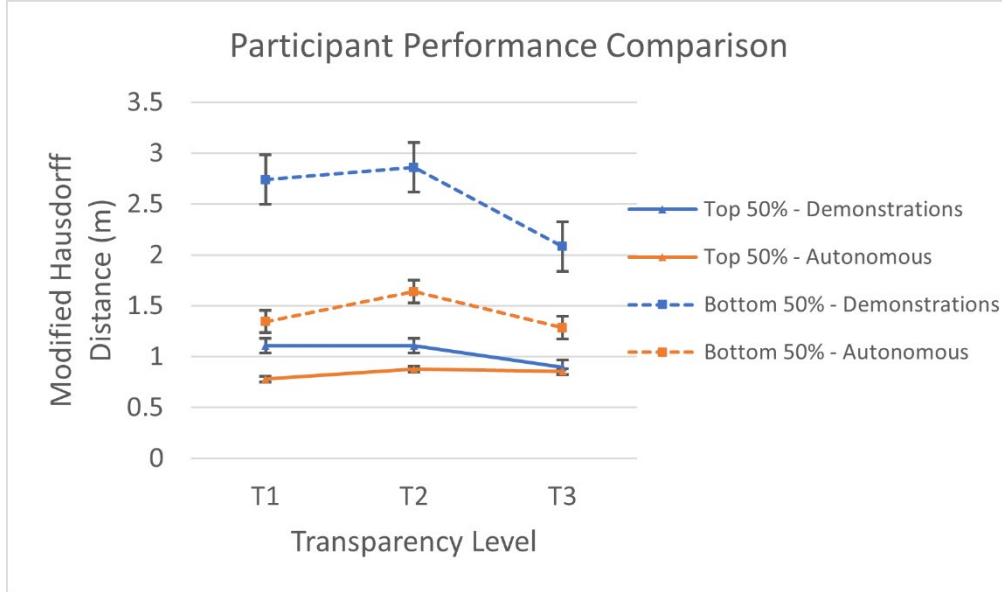


Figure 8. A comparison of the demonstration error and associated autonomous traversal error for the top performing participants compared to the bottom performing participants. Top performance is defined as participants with lowest demonstration trajectory error as computed by the MHD with respect to the ground-truth trajectories. Bars denote SE.

4.2.4 Increased Transparency Reduces Perceived Complexity

LC2 was designed to represent a multi-objective behavior model, that is, the robot needs to avoid dangerous regions and follow the edge of the road. In our second hypothesis, we suggested that the additional complexity in this behavior objective results in more frequent and longer demonstrations than LC1 to achieve the desired performance. Figure 9 breaks down the average number of demonstrations provided by the participants in the two LCs for each waypoint in the three transparency trials. Note that more demonstrations are provided by the LC2 participants than those in LC1 only during the T1 condition. For all other portions of the missions under the other two transparency conditions, the participants from LC1 provide more demonstrations than those under LC2. This supports that

increased transparency also has an impact on reducing the perceived complexity of a problem.

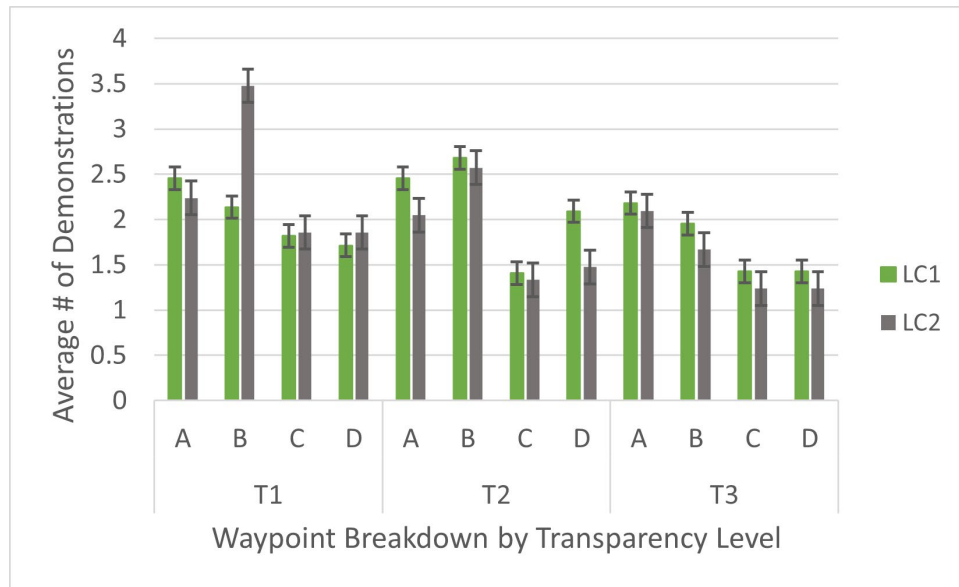


Figure 9. A fine-grained comparison of the average number of demonstrations provided across the life of the mission for the LC1 and LC2 participants. Although LC2 is more complex, fewer demonstrations were provided by the participants in the T2 and T3 cases. Bars denote SE.

We can also break down the demonstrations provided in LC2 to determine what percentage of those were intended to support the first objective of avoiding the dangerous regions and those intended to support the second objective of driving on the edge of the road. Figure 10 shows that as the mission progressed, a smaller percentage of the demonstrations were spent on providing data to learn the objective of avoiding dangerous regions. This indicates that the first objective was learned relatively quickly, and the participants then focused their demonstrations on refining the second objective.

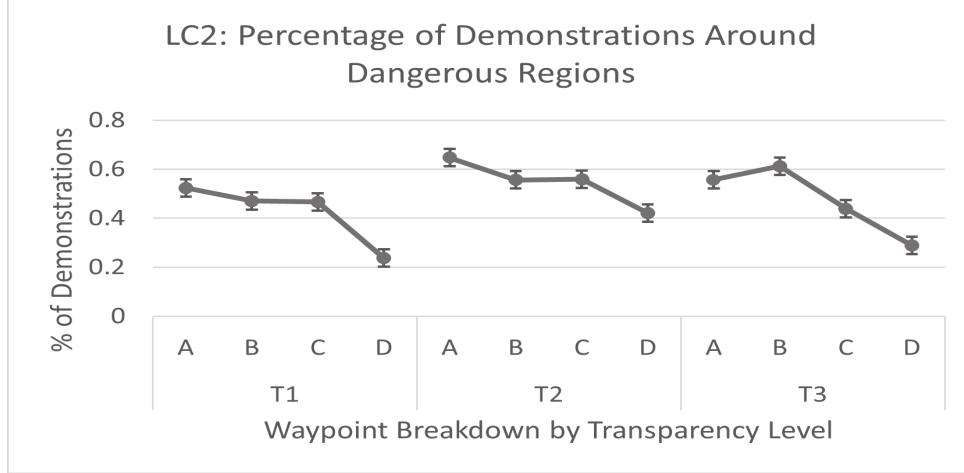


Figure 10. Fine-grained breakdown of the demonstrations provided that were specific to learning to avoid dangerous regions in the LC2 cases. As the mission progresses, fewer of these demonstrations are provided (as seen across all transparency cases), indicating that this portion of the behavior was learned well in the earlier portions of the mission. Bars denote SE.

4.2.5 Transferability of Learned Models to New Missions

Finally, we further evaluate the learned models derived from the participant demonstrations in a distinct area from the simulation trials run during the study. The participants are ranked according to their demonstration quality across the three transparency trials. The final behavior model for each of the top three performing participants is deployed in this new mission that consists of a single waypoint. Under this new mission, we capture the autonomous trajectory planned from the models and again compare it to ground-truth trajectories. This evaluation is designed to showcase that the learning framework does not overfit to specific areas of the simulation, and that these learned models can be used for similar future operational needs.

This evaluation is performed across both LC1 and LC2. Figure 11 provides qualitative results of the LC1 models during this new mission. The first row shows the autonomous trajectories executed using the initial model deployed for the study. This represents the reward function before any human demonstration is provided. Notice that the initial model is designed to favor grass terrain and almost exclusively traverses grass to reach the waypoint. Note that there is also some stochasticity in the trajectories derived from the initial model that is inherent in the global and local planning subsystems of the autonomy stack. Although small differences can be seen across row one, the high-level behavior exhibited from this initial model is stable. The next three rows show the trajectories using models from the top performing participants under each transparency level. As more transparency is introduced, the trajectories seem to match the ground truth more closely. Notice that the autonomous trajectories (Figure 11, blue) tend to be much

closer to the ground truth (Figure 11, cyan) near the beginning of the mission along the left side of the road edge. The deviation from ground truth primarily occurs near the end of the mission when the human-trained models result in trajectories down the middle of the road to reach the waypoint. This indicates that when there is a clear road edge, the learned model captures the behavior well; however, in situations where there are less defined edges, or if following the edges would cause a significant increase in the length of the overall trajectory, then the models tend to simply favor road traversal instead of trying to drive on the edge of the road. Additional empirical evaluation needs to be executed to determine whether more training data could produce a large enough scaling effect to result in the planning system to take longer routes and better mimic the intended behaviors throughout the duration of the mission.

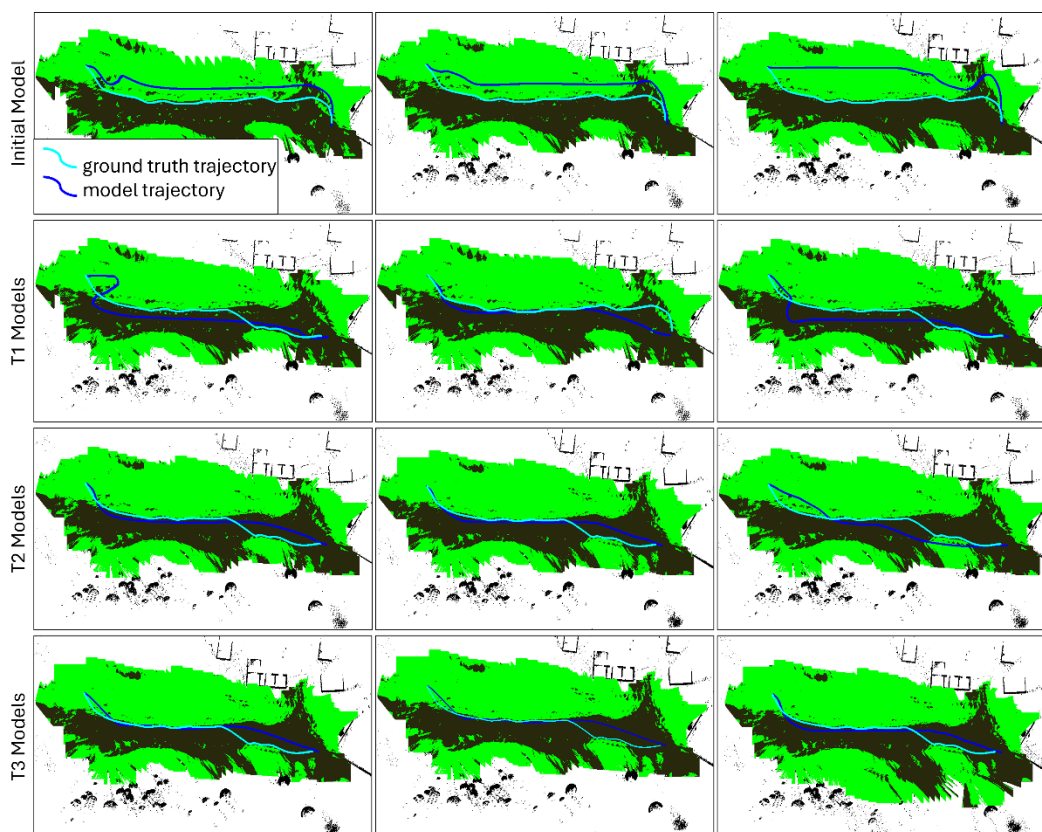


Figure 11. Qualitative trajectory results for several models deployed in a novel LC1 trial compared to those tested in the study. Deployed models are those produced by top performing participants in each of the three transparency levels and the initial behavior model deployed at the start of each trial. Notice that the model trajectories better resemble the ground truth trajectories as more transparency is introduced.

Figure 12 shows the qualitative results for LC2. The first row shows the trajectory results of the initial model used during the study, which represents robot behavior without any human demonstration feedback. In all three runs, the robot plans routes

through the dangerous regions (Figure 12, red circles) because the initial model was not designed to avoid these areas. The following rows all indicate that the human demonstrations provided enough information to teach the robot to effectively avoid these areas.

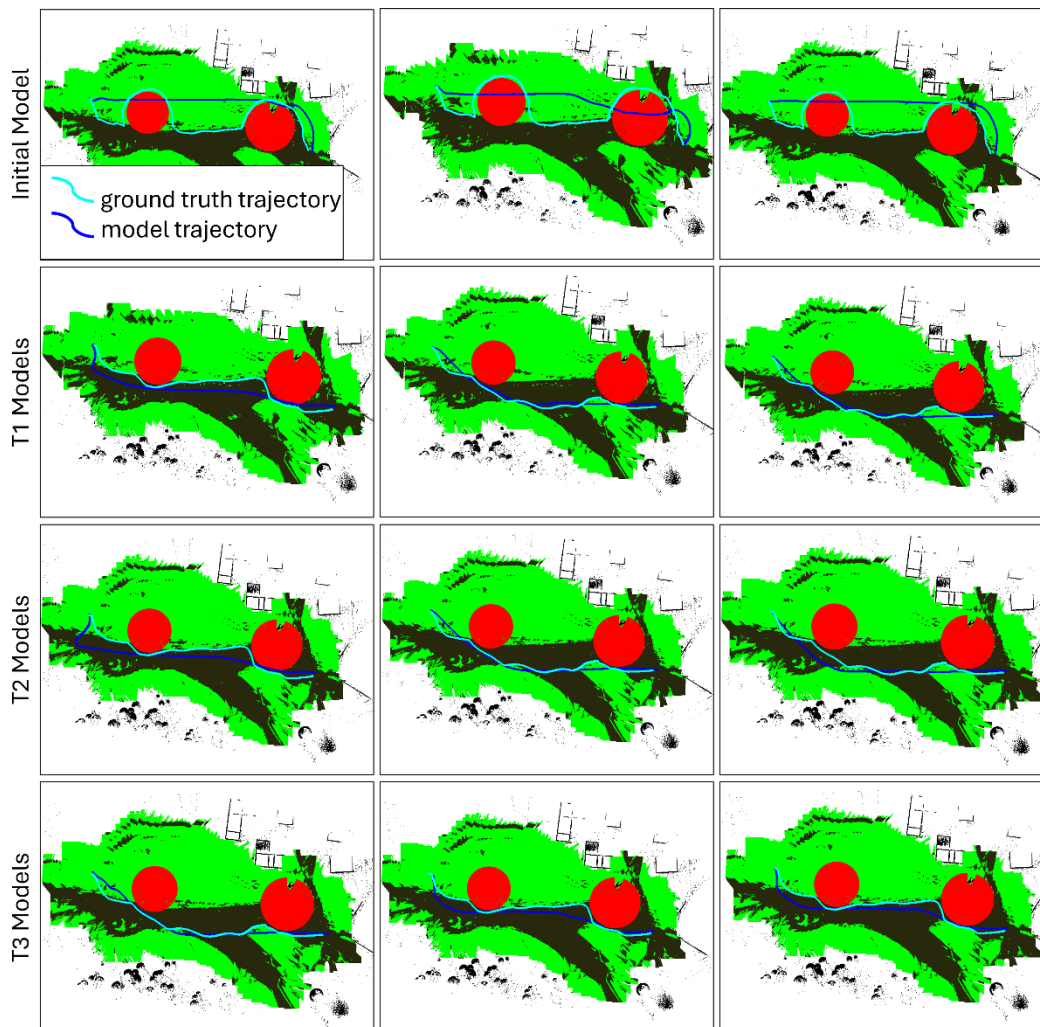


Figure 12. Qualitative trajectory results for several models deployed in a novel LC2 trial compared to those tested in the study. Deployed models are those produced by top performing participants in each of the three transparency levels and the initial behavior model deployed at the start of each trial. Notice that the model trajectories better resemble the ground truth trajectories as more transparency is introduced.

The quantitative results to accompany the qualitative images for this new mission are summarized in Figure 13. These results support the claims that models trained under the T3 transparency trial result in higher quality trajectories, yielding lower errors, with respect to the ground truth.

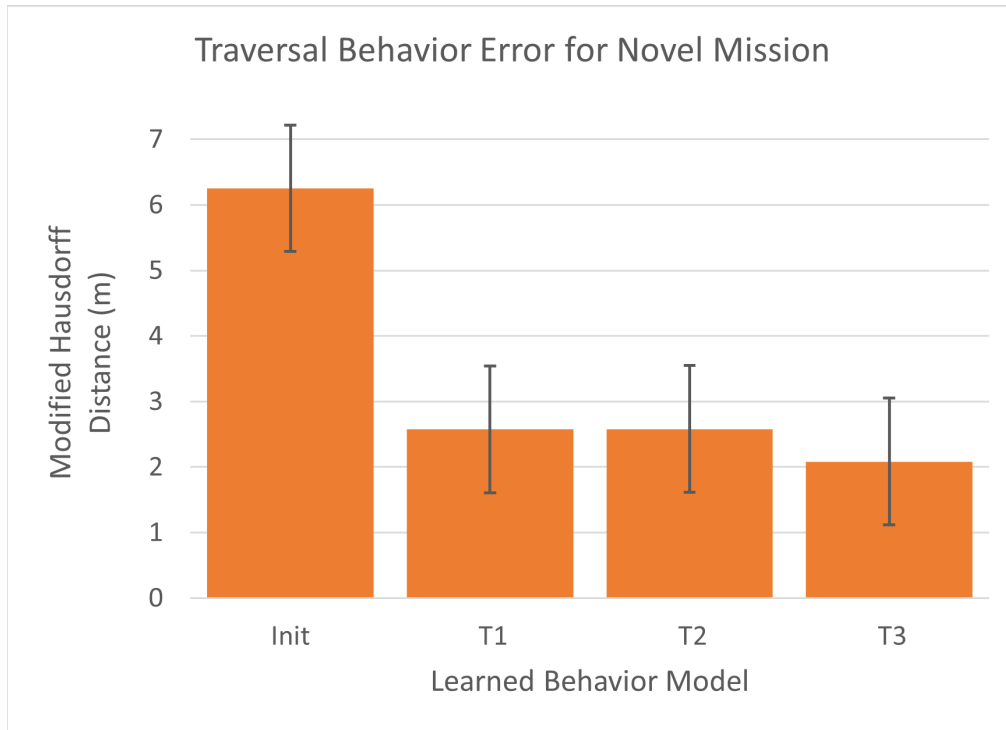


Figure 13. Quantitative measures of the traversal error for the novel mission introduced. Bars denote SE.

4.2.6 Performance Measures Regression Analyses

A supplementary analysis was performed to determine if the performance measures (i.e., the demonstration and autonomy behavior) would predict how participants answered the subjective questionnaires for each trial. This exploratory examination was conducted primarily to inform the design of future studies and is not included in the discussion of findings. Multiple regressions were used to examine these relationships. Duration and distance failed the multicollinearity test and were highly correlated; as such, distance was omitted from this analysis. Below is a summary of the results, organized by transparency condition and reported for each questionnaire. See Appendix Section G-1 for regression statistics and correlations.

The following occurred in the baseline transparency condition (T1):

- Longer duration of participant demonstrations predicted:
 - Lower rating of animacy (GQS)
 - Higher rating of demonstration effectiveness and robot learning (RLQ)
 - Higher evaluation of predicted robot future performance (RLQ)
- Better quality of demonstrations (Quality-MHD) predicted:
 - Higher rating of robot learning (RLQ)
- Fewer errors in robot performance (AutoQuality-MHD) predicted:

- Lower effort (NASA-TLX)
- Lower anthropomorphism, animacy, and perceived intelligence ratings (GQS)
- Greater percentage of distance traveled autonomously (AutoDistance) predicted:
 - Higher reported trust in the robot (TIA scale;)
 - Higher rating of robot learning (RLQ)
 - Higher evaluation of predicted robot future performance (RLQ)

The following occurred in the transparent intent condition (T2):

- Longer Duration of participant demonstrations predicted:
 - Greater temporal demand (NASA-TLX)
- Better quality of demonstrations (Quality-MHD) predicted:
 - Higher perceived intelligence rating (GQS)
 - Lower rating of demonstration effectiveness (RLQ)
- Fewer errors in robot performance (AutoQuality-MHD) predicted:
 - Lower effort (NASA-TLX)
 - Higher ratings of anthropomorphism, animacy, and perceived intelligence (GQS)
 - Higher evaluation of predicted robot future performance (RLQ)
- Greater percentage of distance traveled autonomously (AutoDistance) predicted:
 - Higher reported trust in the robot (TIA)
 - Higher rating of robot learning (RLQ)
 - Higher evaluation of predicted robot future performance (RLQ)

The following occurred in the transparent demonstration impact condition (T3):

- Longer duration of participant demonstrations predicted:
 - Lower rating of animacy (GQS)
 - Higher rating of robot learning (RLQ)
- Better quality of demonstrations (Quality-MHD) predicted:
 - Greater performance satisfaction (NASA-TLX)
 - Lower overall workload (NASA-TLX)
- Fewer errors in robot performance (AutoQuality-MHD) predicted:
 - Lower effort and frustration (NASA-TLX)
 - Greater performance satisfaction (NASA-TLX)
 - Higher reported trust in the robot (TIA)
 - Higher rating of robot learning (RLQ)
 - Robot understanding of the demonstration was clearer (RLQ)

- Greater percentage of distance traveled autonomously (AutoDistance) predicted:
 - Robot understanding of the demonstration was clearer (RLQ)
 - Higher evaluation of predicted robot future performance (RLQ)
 - Reduced physical demand (NASA-TLX)
 - Greater performance satisfaction (NASA-TLX)

4.3 Subjective Measures

4.3.1 NASA-TLX

In repeated measures ANOVAs, there were significant differences across trials for all workload subfactors. See Appendix Section G-2 for NASA-TLX descriptive statistics and analysis results. The addition of transparency in each trial significantly reduces participant workload for subfactors such as mental demand (T1 > T2, $p = 0.015$, $d_s = 0.38$; T2 > T3, $p = 0.004$, $d_s = 0.40$), physical demand (T1 > T2, $p = 0.007$, $d_s = 0.27$), effort (T2 > T3, $p = 0.046$, $d_s = 0.34$), and frustration (T2 > T3, $p = 0.024$, $d_s = 0.32$) and improved their reported performance satisfaction (T1 > T2, $p = 0.036$, $d_s = 0.22$) (Figure 14). In temporal demand, there were moderately significant differences between trials (T1 > T2, $p = 0.059$, $d_s = 0.22$; T2 > T3, $p = 0.060$, $d_s = 0.18$). These findings support hypothesis 3b, that is, with increased transparency, participant cognitive workload would be lessened.

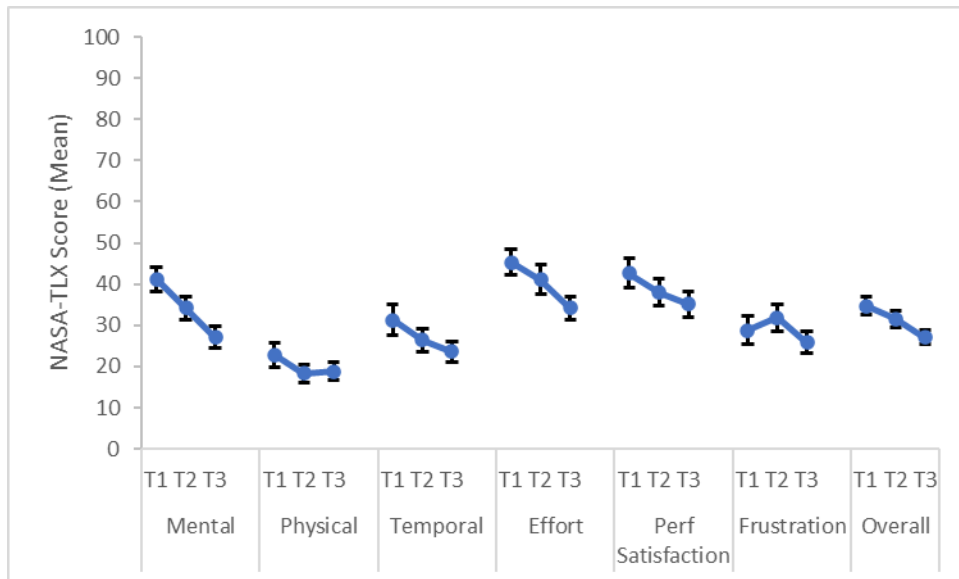


Figure 14. NASA-TLX scores, overall and by subfactor. Bars denote SE.

Recall from Section 4.2.4 that the performance measures seemed to indicate that increased transparency reduces perceived complexity of a problem. To further support this, we also examine differences between transparency levels within each LC using paired t-tests. In LC1, adding intent transparency (T1 – T2) increased participants' frustration (T1 < T2, $p = 0.024$, $d_s = 0.53$) and adding impact transparency (T2 – T3) reduced participants' mental demand (T2 > T3; $p = 0.001$, $d_s = 1.05$), effort ($p = 0.010$, $d_s = 0.72$), and frustration ($p = 0.030$, $d_s = 0.44$). Figure 15 illustrates these findings.

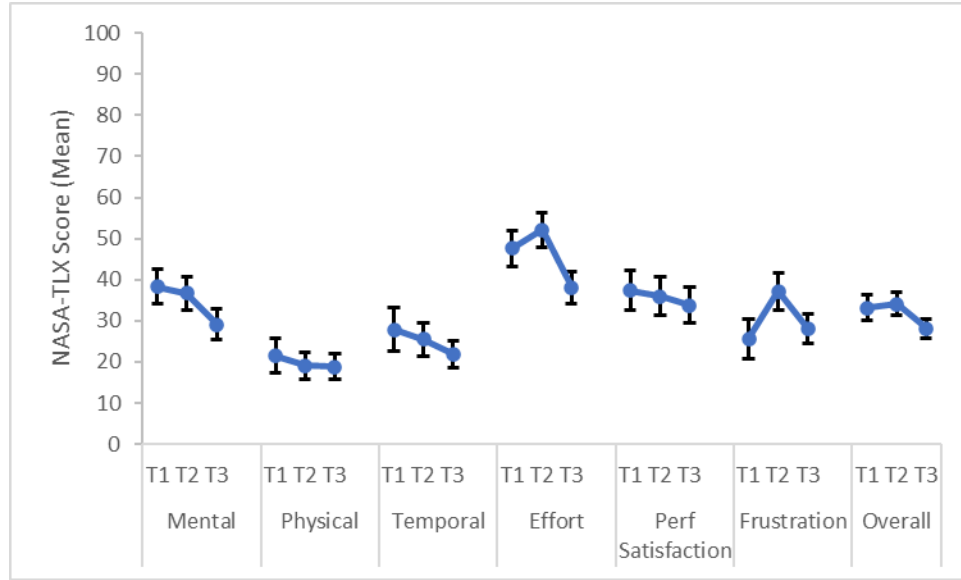


Figure 15. LC1 NASA-TLX scores. Bars denote SE.

In LC2, adding intent transparency (T1 – T2) reduced participant's mental demand (T1 > T2; $p = 0.002$, $d_s = 0.64$), physical demand ($p = 0.019$, $d_s = 0.38$), and effort ($p = 0.023$, $d_s = 0.67$) and increased their performance satisfaction ($p = 0.031$, $d_s = 0.34$), whereas adding impact transparency (T2 – T3) further reduced mental demand ($p = 0.001$, $d_s = 0.83$). Participants' workload was improved by increasing transparency in both conditions. Figure 16 illustrates these trends. These findings show that increased transparency has a stronger positive effect in the LC2 setting compared to LC1, which, again, is the more complex learning task because it has multiple objectives to optimize. Thus, for more complex learning behaviors, increased transparency plays an important role in aiding the human teammate.

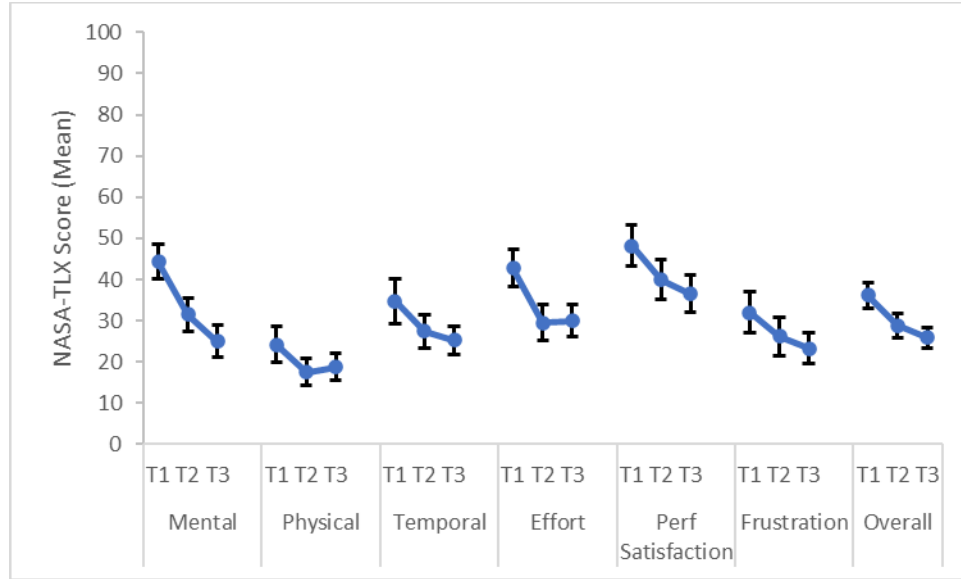


Figure 16. LC2 NASA-TLX scores. Bars denote SE.

4.3.2 Trust in the Robot

Planned comparisons were used to investigate differences in participant trust and distrust in the robot between trials for each LC. See Appendix Section G-3 for the Jian TIA Questionnaire (Jian et al. 2000) descriptive statistics and analysis results. There were no significant differences in overall trust or distrust between transparency. There was a significant difference in trust between T2 and T3 ($T2 < T3$, $p = 0.046$, $d_s = 0.32$) (Figure 17). Although there is a trend for trust in the robot to increase as the transparency level increases, this difference was only significant between T2 and T3. Thus, hypothesis 3c was partially supported.

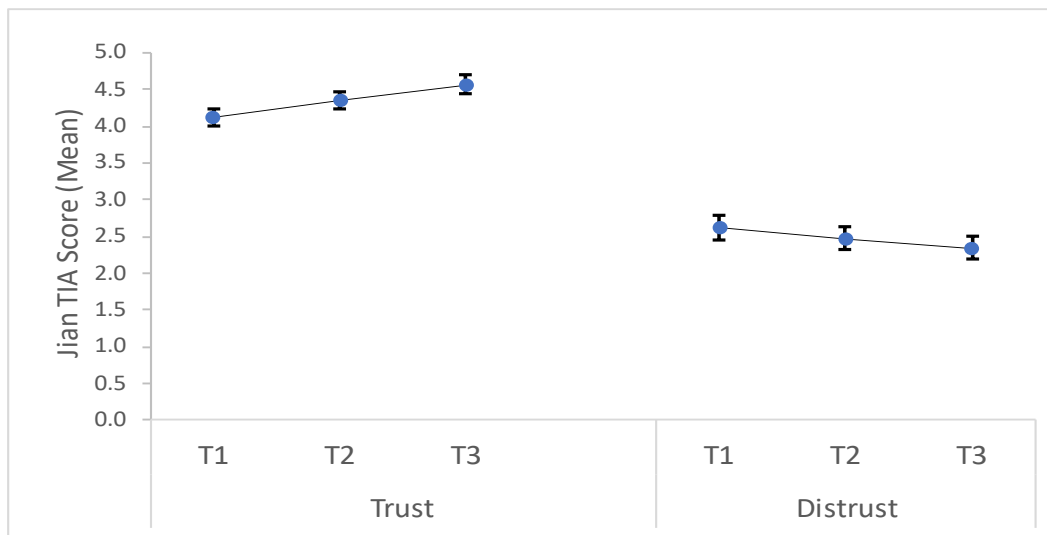


Figure 17. Jian TIA scores. Bars denote SE.

4.3.3 Anthropomorphic Tendencies

Planned comparisons were used to investigate differences in participant scores on the GQS between trials, both overall and for each LC. See Appendix Section G-4 for GQS descriptive statistics and analysis results. Overall, there were few differences in perceptions of the robot between trials. Participants anthropomorphized the robot more in T3 than in T2 ($p = 0.022$, $d_s = 0.37$), indicating that the addition of information about their demonstration resulted in anthropomorphizing the robot more. Participants rated the robot's likeability lower in T2 than in T1 ($p = 0.032$, $d_s = 0.35$), which indicates that adding the robot's intent information had a slight negative effect on their perception of the robot. Except for those minor findings, there was no support for hypothesis 3d, which proposed that adding transparency would increase perceptions of the robot's animacy, intelligence, and likeability (Figure 18).

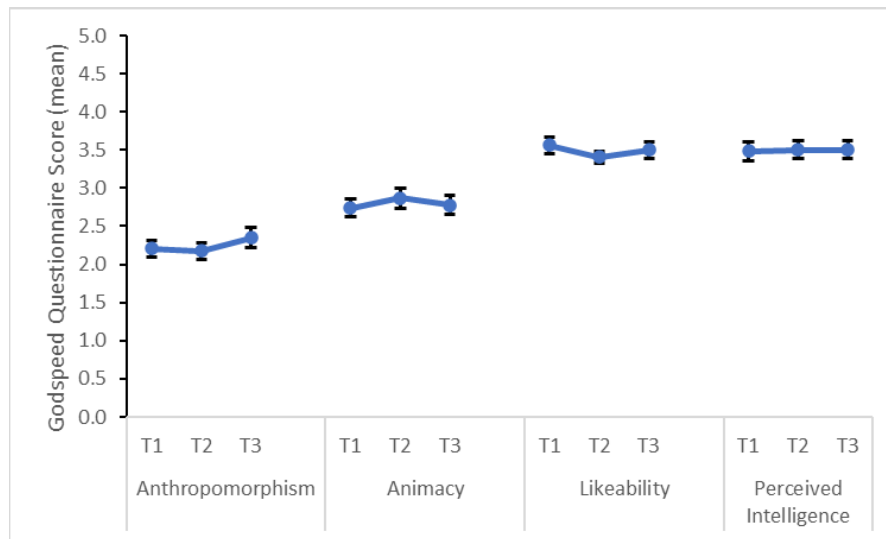


Figure 18. Overall GQS scores, by trial. Bars denote SE.

There were no significant differences in anthropomorphism, animacy, or likeability among transparency conditions in either LC. There was a significant difference in perceived intelligence in LC2 between transparent intent (T2) and impact (T3) ($T2 < T3$, $p = 0.048$, $d_s = 0.14$), wherein the robot was perceived as being more intelligent in the higher transparency trial (T3).

4.3.4 Perception of Robot Learning

At each waypoint, participants answered a short, three-question survey (Appendix E) about their actions and impressions of the robot during that segment. At the final waypoint of each trial, they answered an additional three questions that assessed their projections of future robot behavior. See Appendix Section G-5 for RLQ descriptive statistics and analysis tables.

Overall, there was a significant difference in the number of demonstrations between T1 and T2 ($T1 > T2$, $p = 0.006$, $d_s = 0.48$). There were no other significant differences between trials on any measure (Figure 19).

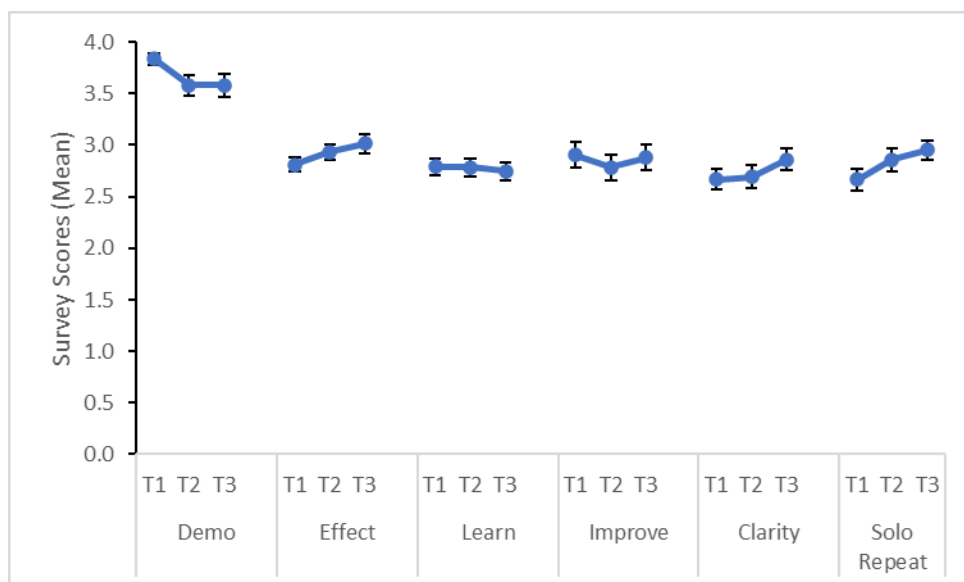


Figure 19. Perception of RLQ scores across transparency conditions (T1, T2, and T3). Bars denote SE.

In LC1, there was a significant difference in the number of demonstrations provided between T1 and T2 ($T1 > T2$, $p = 0.010$, $d_s = 0.72$) and a marginally significant ($p < 0.075$) difference between T2 and T3 ($T2 < T3$, $p = 0.056$, $d_s = 0.50$). The number of demonstrations did not significantly vary in LC2. Participant perception of the effectiveness of their demonstrations increased significantly between T2 and T3 in LC2 ($T2 < T3$, $p = 0.020$, $d_s = 0.41$), but otherwise it did not vary between transparency conditions in either LC. There were no significant differences between trials for perception of robot learning, improvement, clarity, or projections of future performance in either LC (Figure 20).

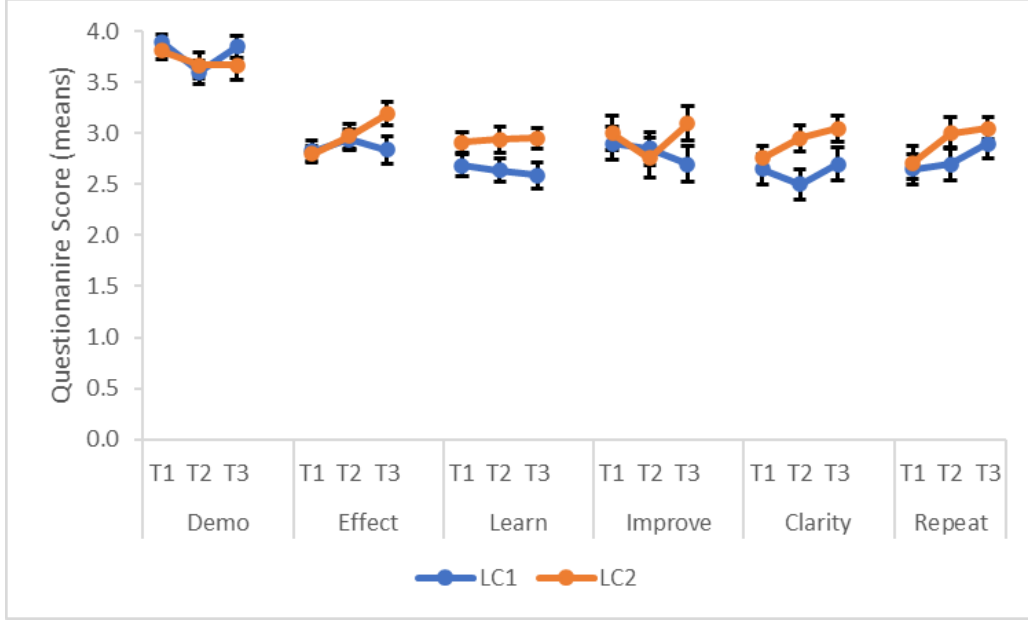


Figure 20. Perception of RLQ scores for LC1 and LC2 across transparency conditions (T1, T2, and T3). Bars denote SE.

5. Discussion

5.1 Synopsis

This study used a high-fidelity robot simulation to investigate online human–robot teaming to achieve adaptable autonomous navigation behavior for the robot. We provide a short discussion on how the data collected in this study helped meet our goals to better understand 1) how ML can be used to facilitate online learning for robot teammates, 2) the impact transparency played in robot learning, and 3) the impact transparency played in human workload, trust, and attributions of the robot.

Overall results from the performance metrics indicate that robot behavior was indeed adapting given human demonstration feedback. Most notably, a significant behavior difference was seen in the updated models produced by participants compared to the initial deployed behavior model by qualitatively comparing the trajectories produced by each model (Figures 11 and 12). This showcases the ability to incorporate ML algorithms in real time during deployment scenarios where a human teammate can aid in the learning and adaption process. This ultimately allows the training pipeline to be forward deployed and used in situ to ensure timely autonomous behavior updates instead of pulling a robot from operational deployment to perform offline updates.

The study results also showed that increasing transparency of robot intent (T2) had a significant effect on the participant's perceived workload. Overall, displaying the robot's intent (i.e., intended planned trajectory) reduced the number of demonstrations participants believed they needed to provide, increasing the distance the robot navigated autonomously compared to the baseline condition, and that is reflected in the lower reported mental and physical demand and increased participant performance satisfaction. This research not only demonstrates the benefits of understanding agent reasoning in reducing cognitive workload and improving performance satisfaction (Lyons and Havig 2014; Wright et al. 2017, 2021), but it also highlights the potential of transparency of agent intent to bring similar benefits.

The study data indicated there was no effect of the intent transparency (T2) on trust in the robot (although it did have a positive trend), and there were no other significant differences in participant attributions of the robot or perceptions of the robot learning. Robot likeability decreased with this increased transparency. Future studies will likely need to gather additional data to help explain this outcome, but we provide some discussion on the possible reasons for this result. First, by increasing transparency, the participant is equipped with more information to make judgments about the robot. In cases in which a robot is operating as desired, this additional information is a good thing because it provides additional confirmation of the robot's intelligence. However, in cases in which the robot may not be operating optimally, this information is going to heighten the participant's awareness to any negative associations they have with this teammate. So, although we generally see that more information (i.e., transparency) often leads to changes in participant associations of the robot, the intelligence of the robot is likely going to dictate whether this change is positive or negative.

In this study, the robot intent was represented by a visualization of the planned trajectory for the robot's current location to the next waypoint. However, the underlying path-planning algorithm is using the real-time sensor measurements that are only able to provide information about the environment that is, at most, 10 m in front of the robot. Thus, in instances in which the robot is far away from the next waypoint, the planned path that is beyond 10 m from sensor observation will appear to have poorer quality. This could be one reason for the decreased likeability, and future efforts could clarify to participants how limited the robot's understanding of its surroundings are, to determine whether this impacts the participant's perception of the robot.

The transparency of the effect of participants' demonstrations (namely, T3) had a significant impact, reducing the number of demonstrations and improving their quality. This increased transparency led to lower reported mental and temporal

demand, perceived effort, and frustration. Feedback on tasks, as we know, is effective for positive cognitive effects and is usually intended to improve performance (Hattie and Timperley 2007). In this instance, it is simply feedback on current performance, and it is unclear how it could contribute to the success of future efforts. However, the participants clearly benefitted from the feedback’s cognitive and performance enhancement benefits (Wisniewski et al. 2020). The number and distance of demonstrations were lower for this trial (T3) than the others. However, the time spent in autonomous operation was as much or more than in previous trials. We interpret these actions by the participants as a sign of increasing confidence in the robot’s ability. This position is supported by a concomitant increase in reported trust in the robot and higher evaluations of anthropomorphism, indicating the participant had a closer connection with the robot (Bartneck et al. 2009).

5.2 Limitations

Although evaluation of the data shows important trends, we provide a list of some limitations that we intend to address in future efforts.

1. Longer trials may have resulted in more dramatic results. Although the missions were designed to be long enough for the human to interact with the system significantly, longer missions would have allowed for more time to improve the algorithm and participant perceptions.
2. The underlying robot traversal task exhibited some ambiguity. Specifically, the simulation environment made it difficult in some areas to identify the edge of the road. This made it difficult to define the ground truth and evaluate demonstration and autonomy quality. In addition, in LC2, there were no precise instructions on how to avoid dangerous regions (i.e., the required standoff from the dangerous region could have been interpreted differently by participants).
3. In some subjective survey data, we found that results were not statistically significant but did have medium effect sizes. Inclusion of more participants for future efforts would increase statistical power and potentially make the findings significant.
4. As briefly mentioned, the range of the robot sensors that are used for the underlying perception and autonomy are limited. For expert roboticists, this is well understood, but because the goal of this research is to enable nonexperts to train the robot teammates, a better explanation of this sensing limitation might impact how the human interacts with the robot. Of course, longer range sensing could be incorporated into the underlying robot capabilities and software, including leveraging satellite data to generate

coarse information about the environment to support robot autonomy decision making.

5. The results may be (at least partially) due to practice and familiarity improving the participant’s performance (i.e., learning effects), since the transparency levels were always presented in a specific order. Future studies should counterbalance the transparency conditions to avoid learning effects.

5.3 Future Work

In addition to some of the topics covered in the study limitations, we want to replicate this work to investigate other forms of human feedback to support rapid adaptation of robot traversal behaviors. Although demonstrations provide one signal of information in which the robot is trying to mimic, full demonstrations during operation require a significant amount of attention from the human teammate. Drawing a human teammate’s attention from other tasks to perform demonstrations may not be ideal for all scenarios. Examples of feedback that could be less time consuming or cognitively difficult for a human to provide include binary feedback or numeric ratings of behavior (i.e., was the previous robot action good or bad; or how good was the previous action on a scale of 1–10) or trajectory comparisons (i.e., given two trajectories, which one is preferred). Furthermore, the study in this report focused on positive demonstrations of behavior, but there is a lot that can be learned from mistakes that are made. Our future studies will also focus on how to process human feedback pertaining to corrections (i.e., resetting the robot when it makes a mistake), where the goal is not to mimic the human actions but to use that information to identify how the robot’s previous actions were suboptimal to update the behavior model.

6. Conclusion

The goal of nonexperts who have the ability to train agents to learn new behaviors after deployment is attainable, and this work demonstrates that not only can this training occur with a few interventions and demonstrations of the desired behavior, but also that specific forms of agent transparency improve the outcomes for both the human user and the agent. When the human user understands the agent’s intent, the result is fewer demonstrations of higher quality and cognitive benefits (i.e., reduced workload) for the user. Adding feedback on the user’s demonstrations further reduced demonstrations, whereas increasing their quality reduced the user’s cognitive workload and bolstered the user’s trust and perceptions of the agent.

7. References

- Alom MZ et al. The history began from AlexNet: a comprehensive survey on deep learning approaches. arXiv; 2018 Mar 3. arXiv:1803.01164. <https://doi.org/10.48550/arXiv.1803.01164>
- Barber DJ et al. Field assessment of multimodal communication for dismounted human-robot teams. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. 2015 Sep;59(1):921–925.
- Bartneck C, Kulić D, Croft E, Zoghbi S. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *Int J Soc Robot*. 2009;1(1):71–81.
- Breazeal C, Thomaz AL. Learning from human teachers with socially guided exploration. 2008 IEEE International Conference on Robotics and Automation; 2008 May. IEEE; c2008. p. 3539–3544.
- Chen JY et al. Situation awareness-based agent transparency. Army Research Laboratory (US); 2014 Apr. Report No.: ARL-TR-6905.
- Chen JYC et al. Situation awareness-based agent transparency and human-autonomy teaming effectiveness. *Theor Issues Ergon Sci*. 2018;19(3):259–282. <https://doi.org/10.1080/1463922X.2017.1315750>
- David RA, Nielsen P. Report of the Defense Science Board summer study on autonomy. Defense Science Board. Department of Defense. 2016 June.
- de Visser E et al. Towards a theory of longitudinal trust calibration in human-robot teams. *Int J Soc Robot*. 2019;12:459–478.
- Drnec K, Metcalfe J. Paradigm development for identifying and validating indicators of trust in automation in the operational environment of human automation integration. *Lect Notes Comput Sci*. 2016;9744:157–167. https://www.doi.org/10.1007/978-3-319-39952-2_16
- Goecks VG, Gremillion GM, Lawhern VJ, Valasek J, Waytowich NR. Efficiently combining demonstrations and interventions for safe training of autonomous systems in real-time. Association for the Advancement of Artificial Intelligence. arXiv; 2019 July 17. arXiv:1810.11545. <https://doi.org/10.48550/arXiv.1810.11545>
- Hart SG, Staveland LE. Development of NASA-TLX (task load index): results of empirical and theoretical research. *Adv Psychol*. 1988;139–183. [https://doi.org/10.1016/s0166-4115\(08\)62386-9](https://doi.org/10.1016/s0166-4115(08)62386-9)

- Hattie J, Timperley H. The power of feedback. *Rev Educat Res*. 2007;77(1):81–112.
- Hinds P, Roberts T, Jones H. Whose job is it anyway? A study of human robot interaction in a collaborative task. *Hum Comput Interact*. 2004;19(1):151–181.
- Hoff KA, Bashir M. Trust in automation: integrating empirical evidence on factors that influence trust. *Hum Fact*. 2015;57(3):407–434. doi: 10.1177/0018720814547570
- Jian JY, Bisantz AM, Drury CG. Foundations for an empirically determined scale of trust in automated systems. *Int J Cogn Ergon*. 2000;4(1):53–71.
- Jiang P, Osteen P, Wigness M, Saripalli S. RELLIS-3D dataset: data, benchmarks and analysis. 2021 IEEE International Conference on Robotics and Automation (ICRA); 2021 30 May–5 June; Xi'an, China. IEEE; c2021. p. 1110–1116.
- Kahn G, Abbeel P, Levine S. BADGR: an autonomous self-supervised learning-based navigation system. *IEEE Robot Autom Lett*. 2021a;6(2):1312–1319.
- Kahn G, Abbeel P, Levine S. Land: learning to navigate from disengagements. *IEEE Robot Autom Lett*. 2021b;6(2):1872–1879.
- Lee JD, See KA. Trust in automation: designing for appropriate reliance. *Hum Fact*. 2004;46(1):50–80.
- Luong M, Pham C. Incremental learning for autonomous navigation of mobile robots based on deep reinforcement learning. *J Intell Robot Syst*. 2021;101(1):1.
- Lyons JB. Being transparent about transparency: a model for human-robot interaction. 2013 AAAI Spring Symposium Series; 2013 Mar.
- Lyons JB, Havig PR. Transparency in a human-machine context: approaches for fostering shared awareness/intent. In: Shumaker R, Lackey S, editors. *Virtual, Augmented and Mixed Reality. Designing and Developing Virtual and Augmented Environments*. VAMR 2014. *Lect Notes Comp Sci*. 2014;8525. Springer. https://doi.org/10.1007/978-3-319-07458-0_18
- Lyons JB, Mahoney S, Wynne KT, Roebke MA. Viewing machines as teammates: a qualitative study. In: 2018 AAAI Spring Symposium; 2018 Mar 26–28; Palo Alto, CA. AAAI Press.

- Merritt SM, Heimbaugh H, LaChapell J, Lee D. I trust it, but I don't know why: Effects of implicit attitudes toward automation on trust in an automated system. *Hum Fact*. 2013;55(3):520–534.
- Mueller ST, Piper BJ. The Psychology Experiment Building Language (PEBL) and PEBL test battery. *J Neurosci Methods*. 2014;222:250–259. <https://doi.org/10.1016/j.jneumeth.2013.10.024>
- Mueller ST. PEBL: The Psychology Experiment Building Language (Version 2.1) [Computer experiment programming language]. PEBL Project; 2014 [accessed 2021 Apr]. <http://pebl.sourceforge.net>
- O'Neill T, McNeese N, Barron A, Schelble B. Human–autonomy teaming: a review and analysis of the empirical literature. *Hum Fact*. 2022;64(5):904–938. <https://doi.org/10.1177/0018720820960865>
- Parasuraman R, Riley V. Humans and automation: use, misuse, disuse, abuse. *Hum Fact*. 1997;39(2):230–253.
- Parasuraman R, Sheridan TB, Wickens CD. A model for types and levels of human interaction with automation. *IEEE Tran Syst Man Cybern A: Sys Hum*. 2000;30(3):286–297.
- Radford A. Improving language understanding by generative pre-training; 2018 June 11. <https://openai.com/index/language-unsupervised/>
- Schaefer KE, Chen JYC, Szalma JL, Hancock PA. A meta-analysis of factors influencing the development of trust in automation: implications for understanding autonomy in future systems. *Hum Fact*. 2016;58:377–400.
- Schaefer KE, Straub ER, Chen JY, Putney J, Evans A. Communicating intent to develop shared situation awareness and engender trust in human-agent teams. *Cogn Syst Res*. 2017;46:26–39. <https://doi.org/10.1016/j.cogsys.2017.02.002>
- Selkowitz AR, Lakhmani SG, Chen JY. Using agent transparency to support situation awareness of the Autonomous Squad Member. *Cogn Syst Res*. 2017;46:13–25.
- Selkowitz AR, Larios CA, Lakhmani SG, Chen JY. Displaying information to support transparency for autonomous platforms. *Advances in Human Factors in Robots and Unmanned Systems. Proceedings of the AHFE 2016 International Conference on Human Factors in Robots and Unmanned Systems*; 2016 July 27–31; Walt Disney World, FL. Springer International Publishing. c2016. p. 161–173.

- Wickens CD. Multiple resources and performance prediction. *Theor Iss Ergon Sci.* 2002;3(2):159–177.
- Wigness M, Rogers JG, Navarro-Serment LE. Robot navigation from human demonstration: Learning control behaviors. *International Conference on Robotics and Automation*; 2018 May 21–26; Brisbane, Australia.
- Wisniewski B, Zierer K, Hattie J. The power of feedback revisited: a meta-analysis of educational feedback research. *Front Psychol.* 2020;10:487662.
- Wooldridge M, Jennings NR. Intelligent agents: theory and practice. *Knowl Eng Rev.* 1995;10(2):115–52.
- Wright JL, Chen JYC, Barnes MJ, Hancock PA. Agent reasoning transparency: the influence of information level on automation-induced complacency. *Army Research Laboratory (US)*; 2017. Report No.: ARL-TR-8044.
- Wright JL, Fiore SM, Foldes J, Sukthankar G. Re-examining the Trust in Autonomy (TIA) scale as trust and distrust factors. *DEVCOM Army Research Laboratory (US)*; 2024. Report No.: ARL-TR-9875.
- Wright JL, Lee J, Schreck JA. Human-autonomy teaming with learning capable agents: performance and workload outcomes. *International Conference on Applied Human Factors and Ergonomics*; 2021 July; Springer; p. 3–10.
- Wynne KT, Lyons JB. An integrative model of autonomous agent teammate-likeness. *Theor Issues Ergon Sci.* 2018;19(3):353–374.
- Xiao X, Liu B, Warnell G, Stone P. Motion planning and control for mobile robot navigation using machine learning: a survey. *Auton Robot.* 2022a;46(5):569–597.
- Xiao X et al. Appl: adaptive planner parameter learning. *Robot Auton Syst.* 2022b;154:104132.

Appendix A. Recruitment Materials

You are invited to participate in a research study.

This study will be conducted at the following locations:

Building 433, 459 or 520, (behind gate 13), Conference Room, Aberdeen Proving Ground, MD

Building 202, Room 4G039, ALC, Adelphi, MD

The purpose of this study is to investigate the impact of human demonstration on machine learning, and the influence of agent transparency on the human. This study will take about 2 hours to complete.

Please ensure you have at least 2 hours available to participate.

When you arrive at the experiment location, you will complete a demographics questionnaire. Then you will complete a training session that includes a practice trial. Participants must demonstrate a minimum level of competence during training and practice to continue with the study. During the session, you will observe a robot traverse a path, and demonstrate for the robot the behavior you wish for the robot to learn (e.g. follow the road, obstacle avoidance) as needed. You will repeat this process several times during the session. After each session is over, you will complete a few questionnaires assessing your experience.

Participation is voluntary. Participants will not be paid.

The U.S. Army Research Laboratory is conducting this study. This research study (ARL 19-067) was reviewed by the ARL HRPP.

If you are interested in participating, please contact:

Maggie B. Wigness, PhD, maggie.b.wigness.civ@army.mil

to reserve your day/time at your preferred location. You will then receive further instructions regarding COVID safety protocols and where to meet research personnel at your scheduled day/time.

Appendix B. Demographics Questionnaire

Demographics Questionnaire

1. General Information

- a. Age: _____ Gender: M F U Handedness: L R
- b. Do you have color deficient vision [i.e., color blindness] Yes No
 - i. If so, what type: Red/Green Blue/Yellow Undefined
- c. Do you have corrected vision (Circle one)? No Glasses Contact Lenses
- d. If so, are you wearing them today? Yes No
- e. How many hours of sleep did you get last night? _____ hours

2. Military Experience

- a. Are you currently in the military? Yes No
 - i. If so, what is your rank? _____
 - ii. What is your job/classification? _____
 - iii. Do you operate robots and/or drones as part of your regular duties? No Yes
 - iv. Have you operated robots and/or drones as part of your regular duties in the past? No Yes
- b. Do you have prior military service? NO YES How long _____
 - i. Did you operate robots and/or drones as part of your regular duties? No Yes

3. AI/ML Experience

- a. What is your level of experience with machine learning and artificial intelligence?
 - i. None
 - ii. Familiar with terms and concepts
 - iii. Experienced - Some understanding of ML methods
 - iv. Expert - Work with AI/ML in robotics
 - v. Expert - Work with AI/ML in other applications

Appendix C. NASA-TLX

NASA-TLX Questionnaire

Please rate your overall impression of demands imposed on you during the exercise.

1. Mental Demand: How much mental and perceptual activity was required (e.g., thinking, looking, searching, etc.)? Was the task easy or demanding, simple or complex, exacting or forgiving?

LOW |---|---|---|---|---|---|---|---| HIGH

1 2 3 4 5 6 7 8 9 10

2. Physical Demand: How much physical activity was required (e.g., pushing, pulling, turning, controlling, activating, etc.)? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious?

LOW |---|---|---|---|---|---|---|---| HIGH

1 2 3 4 5 6 7 8 9 10

3. Temporal Demand: How much time pressure did you feel due to the rate or pace at which the task or task elements occurred? Was the pace slow and leisurely or rapid and frantic?

LOW |---|---|---|---|---|---|---|---| HIGH

1 2 3 4 5 6 7 8 9 10

4. Level of Effort: How hard did you have to work (mentally and physically) to accomplish your level of performance?

LOW |---|---|---|---|---|---|---|---| HIGH

1 2 3 4 5 6 7 8 9 10

5. Level of Frustration: How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task?

LOW |---|---|---|---|---|---|---|---| HIGH

1 2 3 4 5 6 7 8 9 10

6. Performance: How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)? How satisfied were you with your performance in accomplishing these goals?

GOOD |---|---|---|---|---|---|---|---| POOR

1 2 3 4 5 6 7 8 9 10

Appendix D. Trust in Automation

Posttest Trust Survey

Below are statements evaluating your feelings about the robot you were training.
Please circle the number that best describes your feeling or impression.

1 = not at all

7 = extremely

1.	The robot is deceptive.	1	2	3	4	5	6	7
2.	The robot behaved in an underhanded manner.	1	2	3	4	5	6	7
3.	I am suspicious of the intent, action, or outputs of the robot.	1	2	3	4	5	6	7
4.	I am wary of the robot.	1	2	3	4	5	6	7
5.	The actions of the robot will have harmful or injurious outcomes.	1	2	3	4	5	6	7
6.	I am confident in the robot.	1	2	3	4	5	6	7
7.	The robot provides security.	1	2	3	4	5	6	7
8.	The robot has integrity.	1	2	3	4	5	6	7
9.	The robot is dependable.	1	2	3	4	5	6	7
10.	The robot is reliable.	1	2	3	4	5	6	7
11.	I can trust the robot.	1	2	3	4	5	6	7
12.	I am familiar with the robot.	1	2	3	4	5	6	7

The Trust Survey is based on the questionnaire of Human-Computer Trust from Jian et al. (1998)

Appendix E. Godspeed Questionnaire Series

The Godspeed Survey items were presented in random order.

GODSPEED I: ANTHROPOMORPHISM

Please rate your impression of the robot on these scales:

Fake	1	2	3	4	5	Natural
Machinelike	1	2	3	4	5	Humanlike
Unconscious	1	2	3	4	5	Conscious
Artificial	1	2	3	4	5	Lifelike
Moving rigidly	1	2	3	4	5	Moving elegantly

GODSPEED II: ANIMACY

Please rate your impression of the robot on these scales:

Dead	1	2	3	4	5	Alive
Stagnant	1	2	3	4	5	Lively
Mechanical	1	2	3	4	5	Organic
Artificial	1	2	3	4	5	Lifelike
Inert	1	2	3	4	5	Interactive
Apathetic	1	2	3	4	5	Responsive

GODSPEED III: LIKEABILITY

Please rate your impression of the robot on these scales:

Dislike	1	2	3	4	5	Like
Unfriendly	1	2	3	4	5	Friendly
Unkind	1	2	3	4	5	Kind
Unpleasant	1	2	3	4	5	Pleasant
Awful	1	2	3	4	5	Nice

GODSPEED IV: PERCEIVED INTELLIGENCE

Please rate your impression of the robot on these scales:

Incompetent	1	2	3	4	5	Competent
Ignorant	1	2	3	4	5	Knowledgeable
Irresponsible	1	2	3	4	5	Responsible
Unintelligent	1	2	3	4	5	Intelligent
Foolish	1	2	3	4	5	Sensible

GODSPEED V: PERCEIVED SAFETY

Please rate your emotional state on these scales:

Anxious	1	2	3	4	5	Relaxed
Agitated	1	2	3	4	5	Calm
Quiescent	1	2	3	4	5	Surprised

Appendix F. Robot Learning Questionnaire

Robot Learning Evaluation

(Questions 1–6 are presented at the end of each trial [at waypoint 5])

1. Did you demonstrate a desired behavior for the robot during this interval?

No Yes

2. If yes, how effective was your last demonstration?

a – not at all b – somewhat c – mostly d – completely

3. How well has the robot learned the desired behavior?

a – not at all b – somewhat c – mostly d – completely

4. Would additional demonstrations continue to improve robot behavior?

a – not at all b – somewhat c – mostly d – completely

5. How clear was the robot’s understanding of the demonstration intent?

a – not at all b – somewhat c – mostly d – completely

6. Based on the robot’s performance, how well would the robot traverse a similar new path unassisted by human supervisor?

a – Would not successfully navigate the desired path

b – Would successfully navigate the desired path, but with severe deviations.

c – Would successfully navigate the desired path, but with multiple minor deviations.

d – Would successfully navigate the desired path with minimal or no deviations.

(Questions 7–9 are presented at waypoints 2, 3, and 4)

7. Did you demonstrate a desired behavior for the robot during this interval?

No Yes

8. If yes, how effective was your last demonstration?

a – not at all b – somewhat c – mostly d – completely

9. How well has the robot learned the desired behavior?

a – not at all b – somewhat c – mostly d – completely

Appendix G. Analysis Results Tables

G-1 Performance Results

Table G-1. Standard multiple regression of performance outcome variables on NASA-TLX scores.

NASA-TLX factor	Trial	Measure	R^2	ADJ R^2	β	t-test	p-value	Var Exp
Mental demand	T1	Duration	0.043	-0.063	0.03	0.10	0.919	0.0%
		Quality			-0.08	-0.50	0.622	0.7%
		AutoQuality			0.16	0.93	0.357	2.3%
		AutoDistance			-0.08	-0.29	0.772	0.2%
	T2	Duration	0.113	0.015	0.17	0.84	0.406	1.7%
		Quality			-0.10	-0.63	0.534	1.0%
		AutoQuality			0.17	1.06	0.298	2.7%
		AutoDistance			-0.12	-0.61	0.543	0.9%
	T3	Duration	0.106	0.006	0.03	0.14	0.891	0.0%
		Quality			0.16	0.95	0.348	2.2%
		AutoQuality			0.11	0.62	0.536	1.0%
		AutoDistance			-0.25	-1.13	0.267	3.2%
Physical demand	T1	Duration	0.027	-0.081	0.25	0.92	0.362	2.3%
		Quality			-0.05	-0.29	0.774	0.2%
		AutoQuality			-0.09	-0.50	0.623	0.7%
		AutoDistance			0.18	0.66	0.512	1.2%
	T2	Duration	0.116	0.018	0.17	0.88	0.387	1.9%
		Quality			-0.14	-0.87	0.388	1.9%
		AutoQuality			0.05	0.34	0.740	0.3%
		AutoDistance			-0.15	-0.76	0.455	1.4%
	T3	Duration	0.319	0.243	-0.15	-0.74	0.466	1.0%
		Quality			0.25	1.73	0.093	5.6%
		AutoQuality			0.18	1.12	0.269	2.4%
		AutoDistance			-0.58	-2.99	0.005	16.9%
Temporal demand	T1	Duration	0.014	-0.096	0.13	0.48	0.637	0.6%
		Quality			-0.03	-0.20	0.846	0.1%
		AutoQuality			0.00	-0.01	0.994	0.0%
		AutoDistance			0.17	0.64	0.526	1.1%
	T2	Duration	0.145	0.050	0.44	2.27	0.029	12.3%
		Quality			-0.12	-0.74	0.461	1.3%
		AutoQuality			-0.11	-0.71	0.481	1.2%
		AutoDistance			0.23	1.17	0.250	3.2%
	T3	Duration	0.103	0.004	0.11	0.47	0.640	0.6%
		Quality			0.28	1.70	0.098	7.2%
		AutoQuality			0.06	0.35	0.727	0.3%
		AutoDistance			-0.07	-0.33	0.745	0.3%
Effort	T1	Duration	0.091	-0.010	-0.29	-1.09	0.284	3.0%
		Quality			-0.03	-0.16	0.870	0.1%
		AutoQuality			0.30	1.78	0.083	8.0%

Notes: ADJ = adjusted; TLX = Task Load Index; Var Exp = variance explained

Table G-1. Standard multiple regression of performance outcome variables on NASA-TLX scores (continued).

NASA-TLX factor	Trial	Measure	R^2	ADJ R^2	β	t-test	p-value	Var Exp
Effort	T1	Duration	0.091	-0.010	-0.29	-1.09	0.284	3.0%
		Quality			-0.03	-0.16	0.870	0.1%
		AutoQuality			0.30	1.78	0.083	8.0%
		AutoDistance			-0.29	-1.13	0.267	3.2%
	T2	Duration	0.107	0.008	-0.07	-0.36	0.719	0.3%
		Quality			-0.04	-0.23	0.818	0.1%
		AutoQuality			0.28	1.71	0.096	7.2%
		AutoDistance			-0.23	-1.13	0.264	3.2%
	T3	Duration	0.128	0.031	-0.40	-1.71	0.096	7.1%
		Quality			0.04	0.26	0.797	0.2%
		AutoQuality			0.33	1.79	0.081	7.8%
		AutoDistance			-0.17	-0.78	0.441	1.5%
Performance satisfaction	T1	Duration	0.128	0.031	-0.32	-1.23	0.228	3.6%
		Quality			-0.17	-1.10	0.279	2.9%
		AutoQuality			0.27	1.63	0.112	6.4%
		AutoDistance			-0.41	-1.62	0.114	6.4%
	T2	Duration	0.040	-0.067	0.00	-0.02	0.984	0.0%
		Quality			-0.03	-0.15	0.882	0.1%
		AutoQuality			0.18	1.07	0.292	3.0%
		AutoDistance			-0.09	-0.45	0.654	0.5%
	T3	Duration	0.319	0.243	-0.16	-0.76	0.453	1.1%
		Quality			0.39	2.76	0.009	14.4%
		AutoQuality			0.22	1.35	0.187	3.4%
		AutoDistance			-0.44	-2.27	0.029	9.8%
Frustration	T1	Duration	0.043	-0.063	-0.23	-0.85	0.399	1.9%
		Quality			-0.08	-0.47	0.643	0.6%
		AutoQuality			0.17	1.00	0.323	2.7%
		AutoDistance			-0.24	-0.91	0.367	2.2%
	T2	Duration	0.143	0.048	0.27	1.38	0.177	4.5%
		Quality			0.04	0.24	0.810	0.1%
		AutoQuality			0.19	1.17	0.248	3.3%
		AutoDistance			-0.05	-0.27	0.786	0.2%
	T3	Duration	0.107	0.008	-0.18	-0.77	0.448	1.5%
		Quality			0.03	0.20	0.840	0.1%
		AutoQuality			0.35	1.91	0.064	9.1%
		AutoDistance			-0.25	-1.14	0.264	3.2%

Notes: ADJ = adjusted; TLX = Task Load Index; Var Exp = variance explained

Table G-2. Standard multiple regression of performance outcome variables on TIA survey scores.

TIA	Trial	Measure	R^2	ADJ R^2	β	t-test	p-value	Var Exp
Distrust	T1	Duration	0.056	-0.048	-0.28	-1.05	0.303	2.9%
		Quality			-0.07	-0.42	0.675	0.5%
		AutoQuality			0.05	0.31	0.759	0.3%
		AutoDistance			-0.38	-1.41	0.167	5.2%
	T2	Duration	0.012	-0.098	0.07	0.33	0.745	0.3%
		Quality			-0.05	-0.27	0.786	0.2%
		AutoQuality			0.08	0.48	0.635	0.6%
		AutoDistance			0.04	0.21	0.838	0.1%
	T3	Duration	0.090	-0.011	-0.21	-0.88	0.387	1.9%
		Quality			0.14	0.88	0.387	1.9%
		AutoQuality			0.25	1.37	0.179	4.7%
		AutoDistance			-0.23	-1.03	0.308	2.7%
Trust	T1	Duration	0.119	0.021	0.49	1.85	0.072	8.4%
		Quality			0.03	0.21	0.835	0.1%
		AutoQuality			-0.06	-0.34	0.734	0.3%
		AutoDistance			0.56	2.16	0.037	11.4%
	T2	Duration	0.177	0.086	0.10	0.51	0.615	0.6%
		Quality			-0.03	-0.17	0.865	0.1%
		AutoQuality			-0.20	-1.25	0.219	3.6%
		AutoDistance			0.43	2.27	0.029	11.8%
	T3	Duration	0.142	0.046	0.41	1.78	0.084	7.5%
		Quality			0.13	0.83	0.414	1.6%
		AutoQuality			-0.41	-2.28	0.029	12.3%
		AutoDistance			0.36	1.65	0.107	6.5%

Notes: ADJ = adjusted; TIA = Trust in Automation; Var Exp = variance explained

Table G-3. Standard multiple regression of performance outcome variables on Godspeed series survey scores.

GQS survey response	Trial	Measure	R^2	ADJ R^2	β	t-test	p-value	Var Exp
Anthropomorphism	T1	Duration	0.313	0.237	-0.49	-2.14	0.039	8.7%
		Quality			-0.15	-1.10	0.277	2.3%
		AutoQuality			0.51	3.44	<0.001	22.6%
		AutoDistance			-0.19	-0.83	0.410	1.3%
	T2	Duration	0.050	-0.056	-0.16	-0.79	0.433	1.7%
		Quality			-0.12	-0.69	0.492	1.3%
		AutoQuality			0.15	0.88	0.385	2.0%
		AutoDistance			-0.18	-0.88	0.386	2.0%
	T3	Duration	0.090	-0.011	-0.10	-0.43	0.673	0.5%
		Quality			0.26	1.55	0.131	6.0%
		AutoQuality			0.00	0.00	0.998	0.0%
		AutoDistance			-0.16	-0.72	0.474	1.3%
Animacy	T1	Duration	0.277	0.196	-0.52	-2.18	0.036	9.6%
		Quality			-0.13	-0.90	0.372	1.6%
		AutoQuality			0.28	1.85	0.072	6.9%
		AutoDistance			-0.04	-0.18	0.859	0.1%
	T2	Duration	0.034	-0.073	-0.13	-0.61	0.548	1.0%
		Quality			-0.08	-0.45	0.657	0.5%
		AutoQuality			0.11	0.67	0.509	1.2%
		AutoDistance			-0.17	-0.84	0.406	1.9%
	T3	Duration	0.040	-0.066	-0.18	-0.76	0.454	1.5%
		Quality			0.09	0.52	0.604	0.7%
		AutoQuality			-0.03	-0.17	0.863	0.1%
		AutoDistance			-0.13	-0.56	0.580	0.8%
Likeability	T1	Duration	0.066	-0.038	-0.22	-0.84	0.409	1.8%
		Quality			-0.16	-0.99	0.330	2.5%
		AutoQuality			0.20	1.18	0.245	3.6%
		AutoDistance			-0.17	-0.65	0.523	1.1%
	T2	Duration	0.112	0.013	0.00	0.02	0.983	0.0%
		Quality			-0.17	-1.07	0.293	2.8%
		AutoQuality			-0.04	-0.25	0.807	0.1%
		AutoDistance			-0.24	-1.23	0.225	3.8%
	T3	Duration	0.072	-0.032	-0.14	-0.58	0.565	0.9%
		Quality			0.01	0.03	0.972	0.0%
		AutoQuality			-0.01	-0.06	0.956	0.0%
		AutoDistance			-0.32	-1.43	0.161	5.3%
Perceived intelligence	T1	Duration	0.212	0.125	-0.06	-0.23	0.822	0.1%
		Quality			0.02	0.11	0.911	0.0%
		AutoQuality			0.42	2.68	0.011	15.7%
		AutoDistance			0.15	0.62	0.537	0.8%
	T2	Duration	0.164	0.071	-0.27	-1.41	0.166	4.6%
		Quality			-0.33	-2.08	0.045	10.0%
		AutoQuality			0.06	0.38	0.709	0.3%
		AutoDistance			-0.22	-1.15	0.258	3.1%
	T3	Duration	0.093	-0.008	0.16	0.66	0.514	1.1%
		Quality			-0.07	-0.41	0.682	0.4%
		AutoQuality			-0.30	-1.64	0.109	6.8%
		AutoDistance			0.25	1.10	0.277	3.1%

Notes: ADJ = adjusted; GQS = Godspeed Questionnaire Series; Var Exp = variance explained

Table G-4. Standard multiple regression of performance outcome variables on robot learning questionnaire scores.

Survey query	Trial	Measure	R^2	ADJ R^2	β	t-test	p-value	Var Exp
Effectiveness	T1	Duration	0.171	0.084	0.56	2.27	0.029	11.3%
		Quality			-0.22	-1.46	0.153	4.6%
		AutoQuality			-0.21	-1.35	0.185	4.0%
		AutoDistance			0.40	1.64	0.109	5.9%
	T2	Duration	0.214	0.132	-0.15	-0.82	0.417	1.4%
		Quality			0.32	2.16	0.037	9.7%
		AutoQuality			-0.07	-0.48	0.636	0.5%
		AutoDistance			0.18	0.97	0.337	2.0%
	T3	Duration	0.064	-0.034	0.09	0.40	0.688	0.4%
		Quality			0.23	1.39	0.171	4.8%
		AutoQuality			-0.17	-0.96	0.345	2.3%
		AutoDistance			0.07	0.32	0.751	0.3%
Robot learning	T1	Duration	0.375	0.309	0.89	4.14	<0.001	28.2%
		Quality			-0.29	-2.27	0.029	8.5%
		AutoQuality			-0.15	-1.12	0.269	2.1%
		AutoDistance			0.70	3.31	0.002	18.1%
	T2	Duration	0.280	0.204	0.28	1.63	0.112	5.0%
		Quality			0.14	0.96	0.344	1.7%
		AutoQuality			-0.21	-1.43	0.161	3.9%
		AutoDistance			0.59	3.37	0.002	21.5%
	T3	Duration	0.139	0.048	0.44	1.94	0.060	8.5%
		Quality			0.02	0.16	0.874	0.1%
		AutoQuality			-0.39	-2.24	0.031	11.4%
		AutoDistance			0.32	1.51	0.139	5.2%
Improvement	T1	Duration	0.080	-0.017	0.06	0.22	0.829	0.1%
		Quality			-0.01	-0.07	0.946	0.0%
		AutoQuality			-0.28	-1.69	0.099	6.9%
		AutoDistance			-0.03	-0.11	0.915	0.0%
	T2	Duration	0.070	-0.028	0.19	0.96	0.341	2.3%
		Quality			-0.09	-0.54	0.595	0.7%
		AutoQuality			-0.05	-0.29	0.771	0.2%
		AutoDistance			0.33	1.65	0.107	6.7%
	T3	Duration	0.044	-0.059	-0.03	-0.11	0.910	0.0%
		Quality			-0.16	-0.98	0.334	2.5%
		AutoQuality			-0.09	-0.48	0.633	0.6%
		AutoDistance			0.08	0.35	0.726	0.3%

Notes: ADJ = adjusted; Var Exp = variance explained

Table G-4. Standard multiple regression of performance outcome variables on robot learning questionnaire scores (continued).

Survey query	Trial	Measure	R^2	ADJ R^2	β	t-test	p-value	Var Exp
Clarity	T1	Duration	0.112	0.019	0.54	2.12	0.041	10.5%
		Quality			-0.06	-0.42	0.674	0.4%
		AutoQuality			-0.17	-1.02	0.316	2.4%
		AutoDistance			0.42	1.67	0.102	6.6%
	T2	Duration	0.123	0.031	0.11	0.57	0.570	0.8%
		Quality			0.20	1.29	0.204	3.9%
		AutoQuality			-0.25	-1.59	0.120	5.8%
		AutoDistance			0.23	1.21	0.233	3.4%
	T3	Duration	0.178	0.090	0.18	0.81	0.421	1.5%
		Quality			0.04	0.25	0.802	0.1%
		AutoQuality			-0.33	-1.87	0.069	7.8%
		AutoDistance			0.47	2.23	0.032	11.0%
Future performance	T1	Duration	0.145	0.055	0.56	2.22	0.033	11.1%
		Quality			-0.02	-0.15	0.879	0.1%
		AutoQuality			-0.04	-0.26	0.800	0.1%
		AutoDistance			0.59	2.41	0.021	13.0%
	T2	Duration	0.372	0.306	0.19	1.20	0.237	2.4%
		Quality			0.07	0.51	0.615	0.4%
		AutoQuality			-0.52	-3.86	<0.001	24.5%
		AutoDistance			0.46	2.83	0.007	13.2%
	T3	Duration	0.233	0.150	0.07	0.34	0.735	0.2%
		Quality			-0.16	-1.05	0.299	2.3%
		AutoQuality			-0.09	-0.52	0.607	0.6%
		AutoDistance			0.50	2.46	0.019	12.6%

Notes: ADJ = adjusted; Var Exp = variance explained

Table G-5. Performance measure correlations with confidence intervals.

Trial	Measures	Pearson's correlation (two-tailed)	Sig. (two-tailed)	95% confidence interval (two-tailed) ^a	
				Lower	Upper
T1	Duration - Distance	0.95	<0.001	0.90	0.97
	Duration - Quality MHD	0.02	0.918	-0.29	0.31
	Duration - AutoQuality MHD	0.19	0.219	-0.12	0.47
	Duration - AutoDistance	-0.77	<0.001	-0.87	-0.61
	Distance - Quality MHD	0.03	0.827	-0.27	0.33
	Distance - AutoQuality MHD	0.18	0.241	-0.12	0.46
	Distance - AutoDistance	-0.78	<0.001	-0.87	-0.62
	Quality MHD - AutoQuality MHD	0.01	0.968	-0.29	0.31
	Quality MHD - AutoDistance	-0.03	0.863	-0.32	0.28
	AutoQuality MHD - AutoDistance	0.03	0.825	-0.27	0.33
T2	Duration - Distance	0.95	<0.001	0.91	0.97
	Duration - Quality MHD	-0.08	0.627	-0.37	0.23
	Duration - AutoQuality MHD	0.14	0.380	-0.17	0.42
	Duration - AutoDistance	-0.59	<0.001	-0.76	-0.35
	Distance - Quality MHD	0.02	0.923	-0.29	0.31
	Distance - AutoQualityMHD	0.14	0.365	-0.17	0.42
	Distance - AutoDistance	-0.61	<0.001	-0.77	-0.38
	Quality MHD - AutoQuality MHD	0.22	0.159	-0.09	0.49
	Quality MHD - AutoDistance	0.19	0.225	-0.12	0.46
	AutoQualityMHD - AutoDistance	0.02	0.916	-0.29	0.32
T3	Duration - Distance	0.97	<0.001	0.95	0.98
	Duration - Quality MHD	-0.17	0.274	-0.45	0.14
	Duration - AutoQuality MHD	0.29	0.061	-0.01	0.54
	Duration - AutoDistance	-0.61	<0.001	-0.77	-0.38
	Distance - Quality MHD	-0.13	0.396	-0.42	0.17
	Distance - AutoQualityMHD	0.29	0.062	-0.01	0.54
	Distance - AutoDistance	-0.61	<0.001	-0.77	-0.38
	Quality MHD - AutoQuality MHD	0.05	0.763	-0.26	0.34
	Quality MHD - AutoDistance	0.01	0.968	-0.29	0.31
	AutoQuality MHD - AutoDistance	0.13	0.411	-0.18	0.41

Notes: MHD = modified Hausdorff distance; Sig. = significance

^a Estimation is based on Fisher's r-to-z transformation

G-2 NASA-TLX Workload Questionnaire Results

Table G-6. Overall descriptive statistics for NASA-TLX questionnaire results.

Measure	Trial	Mean	SD	SE
Mental	T1	41.22	18.93	2.96
	T2	34.15	17.92	2.80
	T3	27.07	17.17	2.68
Physical	T1	22.80	19.17	2.99
	T2	18.29	14.30	2.23
	T3	18.78	14.13	2.21
Temporal	T1	31.22	23.95	3.74
	T2	26.46	18.11	2.83
	T3	23.54	15.22	2.38
Effort	T1	45.24	20.19	3.15
	T2	41.10	22.49	3.51
	T3	34.15	17.82	2.78
Performance satisfaction	T1	42.68	22.81	3.56
	T2	37.93	20.98	3.28
	T3	35.12	20.11	3.14
Frustration	T1	28.78	22.16	3.46
	T2	31.83	21.21	3.31
	T3	25.73	16.87	2.63
Overall workload	T1	34.66	13.81	2.16
	T2	31.50	13.00	2.03
	T3	27.04	11.11	1.74

Notes: SD = standard deviation; SE = standard error; TLX = Task Load Index

Table G-7. Repeated measures ANOVA results for overall NASA-TLX scores between trials.

Measure	<i>F</i>	<i>df</i>	Sig.	η^2_p
Mental	14.10	2, 39	<0.001	0.42
Physical	5.84	1.6, 62.4 ^a	0.009	0.13
Temporal	5.98	1.5, 61.2 ^a	0.008	0.13
Effort	5.28	1.5, 61.1 ^a	0.013	0.12
Performance satisfaction	4.84	2, 39	0.013	0.20
Frustration	2.80	2, 39	0.073	0.13
Overall workload	13.66	1.7, 67.2 ^a	<0.001	0.26

Notes: ANOVA = analysis of variance; Sig. = significance; TLX = Task Load Index

^a Indicates Greenhouse–Geisser correction applied

Table G-8. Post hoc paired t-test results for overall NASA-TLX scores.

Measure	Trials	<i>t</i> (20)	Sig.	Cohen's <i>d_s</i>
Mental	T1–T2	2.54	0.015	0.38
	T2–T3	3.02	0.004	0.40
Physical	T1–T2	2.86	0.007	0.27
	T2–T3	–0.49	0.628	0.03
Temporal	T1–T2	1.95	0.059	0.22
	T2–T3	1.94	0.060	0.18
Effort	T1–T2	0.98	0.333	0.19
	T2–T3	2.06	0.046	0.34
Performance satisfaction	T1–T2	2.17	0.036	0.22
	T2–T3	1.28	0.207	0.14
Frustration	T1–T2	–0.97	0.339	0.14
	T2–T3	2.35	0.024	0.32
Overall workload	T1–T2	1.80	0.080	0.24
	T2–T3	3.39	0.002	0.37

Notes: Sig. = significance; TLX = Task Load Index

Table G-9. NASA-TLX scores descriptive statistics by learning condition (LC).

Measure	Trial	LC1 <i>N</i> = 21			LC2 <i>N</i> = 20		
		Mean	SD	SE	Mean	SD	SE
Mental	T1	38.33	14.52	4.13	44.25	22.67	4.23
	T2	36.67	18.73	3.92	31.50	17.10	4.01
	T3	29.05	16.93	3.77	25.00	17.62	3.86
Physical	T1	21.43	17.62	4.23	24.25	21.04	4.33
	T2	19.05	15.05	3.16	17.50	13.81	3.23
	T3	18.81	14.99	3.12	18.75	13.56	3.20
Temporal	T1	27.86	18.21	5.24	34.75	28.86	5.37
	T2	25.48	17.10	3.99	27.50	19.50	4.09
	T3	21.90	15.53	3.34	25.25	15.09	3.42
Effort	T1	47.62	18.62	4.43	42.75	21.91	4.54
	T2	52.14	21.65	4.28	29.50	17.16	4.38
	T3	38.10	16.84	3.83	30.00	18.28	3.93
Performance satisfaction	T1	37.38	19.53	4.89	48.25	25.09	5.01
	T2	35.95	18.75	4.61	40.00	23.40	4.73
	T3	33.81	18.02	4.43	36.50	22.48	4.54
Frustration	T1	25.71	20.57	4.85	32.00	23.81	4.97
	T2	37.14	22.22	4.53	26.25	19.05	4.64
	T3	28.10	18.47	3.69	23.25	15.07	3.78
Overall workload	T1	33.17	11.53	3.03	36.22	16.01	3.11
	T2	34.03	12.67	2.81	28.85	13.13	2.88
	T3	28.10	10.69	2.44	25.93	11.72	2.50

Notes: SD = standard deviation; SE = standard error; TLX = Task Load Index

Table G-10. NASA-TLX paired t-test results, within learning condition (LC).

Measure	Trials	LC1 N = 21			LC2 N = 20		
		<i>t</i> (20)	Sig.	Cohen's <i>d</i>	<i>t</i> (19)	Sig.	Cohen's <i>d</i>
Mental	T1–T2	0.42	0.680	0.10	3.58	0.002	0.64
	T2–T3	4.88	<0.001	1.05	3.99	<0.001	0.83
Physical	T1–T2	1.36	0.188	0.15	2.57	0.019	0.38
	T2–T3	0.16	0.874	0.02	−0.93	0.367	0.09
Temporal	T1–T2	0.80	0.431	0.13	1.84	0.081	0.29
	T2–T3	1.61	0.122	0.22	1.07	0.297	0.13
Effort	T1–T2	−0.75	0.461	0.22	2.47	0.023	0.67
	T2–T3	2.85	0.010	0.72	−0.12	0.903	0.03
Performance satisfaction	T1–T2	0.57	0.576	0.07	2.33	0.031	0.34
	T2–T3	0.66	0.517	0.12	1.17	0.257	0.15
Frustration	T1–T2	−2.44	0.024	0.53	1.77	0.093	0.27
	T2–T3	2.34	0.030	0.44	0.88	0.390	0.17
Overall workload	T1–T2	−0.36	0.724	0.07	3.28	0.004	0.50
	T2–T3	3.29	0.004	0.51	1.53	0.142	0.23

Notes: Sig. = significance; TLX = Task Load Index

Table G-11. NASA-TLX factor correlations with performance measures.

NASA-TLX factor	Performance measure	Trial	Pearson's <i>r</i>	Sig. (two-tailed)	95% confidence interval (two-tailed) ^a	
					Lower	Upper
Mental demand	Duration	T1	0.12	0.463	−0.20	0.41
		T2	0.27	0.089	−0.04	0.53
		T3	0.19	0.225	−0.12	0.47
	Distance	T1	0.08	0.598	−0.23	0.38
		T2	0.27	0.092	−0.04	0.53
		T3	0.17	0.276	−0.14	0.46
	Quality-MHD	T1	−0.08	0.631	−0.38	0.24
		T2	−0.10	0.535	−0.40	0.21
		T3	0.15	0.336	−0.16	0.44
	AutoQuality-MHD	T1	0.16	0.304	−0.15	0.45
		T2	0.17	0.285	−0.14	0.45
		T3	0.10	0.544	−0.22	0.39
	AutoDistance	T1	−0.09	0.570	−0.39	0.22
		T2	−0.24	0.138	−0.51	0.08
		T3	−0.26	0.107	−0.52	0.06
Physical demand	Duration	T1	0.10	0.540	−0.22	0.39
		T2	0.28	0.077	−0.03	0.54
		T3	0.22	0.175	−0.10	0.49
	Distance	T1	0.03	0.836	−0.28	0.34
		T2	0.21	0.190	−0.11	0.49
		T3	0.17	0.294	−0.15	0.45
	Quality-MHD	T1	−0.05	0.757	−0.35	0.26
		T2	−0.17	0.285	−0.45	0.14
		T3	0.28	0.078	−0.03	0.54
	AutoQuality-MHD	T1	−0.03	0.842	−0.34	0.28
		T2	0.05	0.780	−0.27	0.35
		T3	0.07	0.664	−0.24	0.37

Notes: MHD = modified Hausdorff distance; Sig. = significance; TLX = Task Load Index

^a Estimation is based on Fisher's *r*-to-*z* transformation

Table G-11. NASA-TLX factor correlations with performance measures (continued).

NASA-TLX factor	Performance measure	Trial	Pearson's r	Sig. (two-tailed)	95% confidence interval (two-tailed) ^a	
					Lower	Upper
Physical demand	AutoDistance	T1	-0.02	0.909	-0.32	0.29
		T2	-0.28	0.080	-0.54	0.03
		T3	-0.46	0.002	-0.68	-0.18
Temporal demand	Duration	T1	0.00	0.983	-0.31	0.30
		T2	0.30	0.057	-0.01	0.56
		T3	0.13	0.427	-0.19	0.42
	Distance	T1	-0.02	0.878	-0.33	0.29
		T2	0.27	0.082	-0.04	0.54
		T3	0.11	0.511	-0.21	0.40
	Quality-MHD	T1	-0.04	0.823	-0.34	0.27
		T2	-0.14	0.396	-0.43	0.18
		T3	0.26	0.098	-0.05	0.53
	AutoQuality-MHD	T1	0.03	0.853	-0.28	0.33
		T2	-0.08	0.637	-0.37	0.24
		T3	0.10	0.534	-0.21	0.40
	AutoDistance	T1	0.07	0.647	-0.24	0.37
		T2	-0.06	0.727	-0.36	0.26
		T3	-0.13	0.412	-0.42	0.18
Effort	Duration	T1	0.00	0.977	-0.31	0.30
		T2	0.10	0.524	-0.21	0.40
		T3	-0.20	0.200	-0.48	0.11
	Distance	T1	0.03	0.875	-0.28	0.33
		T2	0.11	0.512	-0.21	0.40
		T3	-0.22	0.173	-0.49	0.10
	Quality-MHD	T1	-0.02	0.902	-0.33	0.29
		T2	-0.01	0.937	-0.32	0.30
		T3	0.12	0.441	-0.19	0.42
	AutoQuality-MHD	T1	0.24	0.137	-0.08	0.51
		T2	0.26	0.103	-0.05	0.52
		T3	0.19	0.236	-0.13	0.47
	AutoDistance	T1	-0.06	0.709	-0.36	0.25
		T2	-0.19	0.245	-0.47	0.13
		T3	0.12	0.473	-0.20	0.41
Performance satisfaction	Duration	T1	0.05	0.759	-0.26	0.35
		T2	0.08	0.629	-0.24	0.38
		T3	0.11	0.490	-0.20	0.40
	Distance	T1	0.04	0.807	-0.27	0.34
		T2	0.05	0.752	-0.26	0.35
		T3	0.09	0.587	-0.23	0.38
	Quality-MHD	T1	-0.16	0.313	-0.45	0.15
		T2	0.00	0.988	-0.31	0.31
		T3	0.43	0.005	0.14	0.65
	AutoQuality-MHD	T1	0.19	0.225	-0.12	0.47
		T2	0.17	0.276	-0.14	0.46
		T3	0.13	0.419	-0.19	0.42

Notes: MHD = modified Hausdorff distance; Sig. = significance; TLX = Task Load Index

^a Estimation is based on Fisher's r-to-z transformation

Table G-11. NASA-TLX factor correlations with performance measures (continued).

NASA-TLX factor	Performance measure	Trial	Pearson's r	Sig. (two-tailed)	95% confidence interval (two-tailed) ^a	
					Lower	Upper
Performance satisfaction	AutoDistance	T1	-0.15	0.338	-0.44	0.16
		T2	-0.09	0.565	-0.39	0.22
		T3	-0.32	0.044	-0.57	-0.01
	Duration	T1	-0.01	0.943	-0.32	0.30
		T2	0.32	0.040	0.02	0.57
		T3	0.07	0.660	-0.24	0.37
	Distance	T1	-0.09	0.560	-0.39	0.22
		T2	0.30	0.061	-0.01	0.55
		T3	0.03	0.867	-0.28	0.33
Frustration	Quality-MHD	T1	-0.07	0.658	-0.37	0.24
		T2	0.05	0.754	-0.26	0.35
		T3	0.08	0.623	-0.23	0.38
	AutoQuality-MHD	T1	0.12	0.453	-0.19	0.41
		T2	0.23	0.142	-0.08	0.50
		T3	0.27	0.093	-0.05	0.53
	AutoDistance	T1	-0.06	0.723	-0.36	0.25
		T2	-0.20	0.208	-0.48	0.11
		T3	-0.10	0.553	-0.39	0.22

Notes: MHD = modified Hausdorff distance; Sig. = significance; TLX = Task Load Index

^a Estimation is based on Fisher's r-to-z transformation

G-3 Jian Trust in Autonomy Questionnaire Results

Table G-12. Overall Jian TIA descriptive statistics.

Measure	Trial	Mean	SD	SE
Trust	T1	4.12	1.10	0.17
	T2	4.36	0.99	0.15
	T3	4.57	1.01	0.16
Distrust	T1	2.63	1.03	0.16
	T2	2.48	1.00	0.16
	T3	2.34	0.97	0.15

Notes: SD = standard deviation; SE = standard error; TIA = Trust in Automation

Table G-13. Repeated measures ANOVA results for Jian TIA questionnaire scores.

Condition	Measure	F	df	Sig.	η^2_p
Overall	Trust	5.53	1.6, 66.0 ^a	0.009	0.12
	Distrust	2.52	1.7, 65.9 ^a	0.122	0.05
LC1	Trust	1.81	2, 18	0.192	0.17
	Distrust	1.83	2, 18	0.189	0.17
LC2	Trust	3.26	1.6, 31.3 ^a	0.063	0.14
	Distrust	0.93	2, 19	0.413	0.09

Notes: ANOVA = analysis of variance; LC = learning condition;

Sig. = significance; TIA = Trust in Automation

^a Indicates Greenhouse–Geisser correction applied

Table G-14. Post hoc paired t-test results for Jian TIA overall scores.

Measure	Trials	<i>t</i> (40)	Sig.	Cohen's <i>d</i> _s
Trust	T1–T2	–1.51	0.138	0.24
	T2–T3	–2.06	0.046	0.32
Distrust	T1–T2	0.94	0.353	0.15
	T2–T3	1.29	0.203	0.20

Notes: Sig. = significance; TIA = Trust in Automation

Table G-15. Jian Trust in Automation descriptive statistics, by learning condition (LC).

Condition	Measure	Trial	Mean	SD	SE
LC1	Distrust	T1	2.61	1.09	0.24
		T2	2.35	0.87	0.19
		T3	2.23	0.89	0.20
	Trust	T1	3.91	1.01	0.23
		T2	4.19	1.08	0.24
		T3	4.30	0.97	0.22
LC2	Distrust	T1	2.65	1.01	0.22
		T2	2.60	1.11	0.24
		T3	2.45	1.05	0.23
	Trust	T1	4.33	1.16	0.25
		T2	4.52	0.89	0.19
		T3	4.82	1.01	0.22

Notes: SD = standard deviation; SE = standard error

Table G-16. Jian TIA paired t-test results within learning condition (LC).

Learning condition	Measure	Trials	<i>t</i> (<i>I</i>)	<i>df</i>	Sig.	Cohen's <i>d</i>
LC1	Distrust	T1–T2	1.02	19	0.320	0.23
		T2–T3	0.75	19	0.462	0.17
	Trust	T1–T2	–1.37	19	0.185	0.31
		T2–T3	–0.76	19	0.458	0.17
LC2	Distrust	T1–T2	0.23	20	0.817	0.05
		T2–T3	1.07	20	0.299	0.23
	Trust	T1–T2	–0.80	20	0.432	0.18
		T2–T3	–2.10	20	0.048	0.46

Notes: Sig. = significance; TIA = Trust in Automation

Table G-17. Jian TIA measures correlations with performance measures.

Jian TIA measure	Performance measure	Trial	Pearson's r	Sig. (two-tailed)	95% confidence interval (two-tailed) ^a	
					Lower	Upper
Overall trust	Duration	T1	0.04	0.820	-0.27	0.34
		T2	-0.17	0.280	-0.46	0.14
		T3	0.01	0.970	-0.30	0.31
	Distance	T1	-0.01	0.937	-0.32	0.30
		T2	-0.14	0.366	-0.43	0.17
		T3	-0.01	0.958	-0.32	0.30
	Quality-MHD	T1	0.05	0.751	-0.26	0.35
		T2	0.01	0.946	-0.30	0.32
		T3	-0.11	0.479	-0.41	0.20
	AutoQuality-MHD	T1	0.05	0.753	-0.26	0.35
		T2	-0.19	0.224	-0.47	0.12
		T3	-0.36	0.022	-0.60	-0.06
	AutoDistance	T1	0.22	0.168	-0.09	0.49
		T2	0.29	0.070	-0.02	0.55
		T3	0.11	0.495	-0.20	0.40
Trust	Duration	T1	0.05	0.774	-0.27	0.35
		T2	-0.18	0.250	-0.47	0.13
		T3	0.05	0.773	-0.27	0.35
	Distance	T1	0.00	0.978	-0.30	0.31
		T2	-0.16	0.315	-0.45	0.15
		T3	0.01	0.945	-0.30	0.32
	Quality-MHD	T1	0.02	0.891	-0.29	0.33
		T2	0.00	0.992	-0.31	0.31
		T3	0.05	0.777	-0.27	0.35
	AutoQuality-MHD	T1	0.06	0.732	-0.26	0.36
		T2	-0.18	0.252	-0.46	0.13
		T3	-0.24	0.138	-0.51	0.08
	AutoDistance	T1	0.18	0.264	-0.14	0.46
		T2	0.37	0.018	0.07	0.61
		T3	0.06	0.726	-0.26	0.36
Distrust	Duration	T1	0.02	0.924	-0.29	0.32
		T2	0.06	0.719	-0.25	0.36
		T3	-0.02	0.920	-0.32	0.29
	Distance	T1	0.04	0.782	-0.27	0.35
		T2	0.03	0.833	-0.28	0.34
		T3	-0.04	0.803	-0.34	0.27
	Quality-MHD	T1	-0.06	0.707	-0.36	0.25
		T2	-0.03	0.869	-0.33	0.28
		T3	0.19	0.232	-0.12	0.47
	AutoQuality-MHD	T1	-0.01	0.931	-0.32	0.29
		T2	0.08	0.608	-0.23	0.38
		T3	0.17	0.291	-0.15	0.45
	AutoDistance	T1	-0.15	0.340	-0.44	0.16
		T2	0.00	0.978	-0.31	0.30
		T3	-0.07	0.660	-0.37	0.24

Notes: MHD = modified Hausdorff distance; TLX = Task Load Index

^a Estimation is based on Fisher's r-to-z transformation

G-4 Godspeed Series Questionnaire Results

Table G-18. Overall Godspeed Questionnaire Series descriptive statistics.

Measure	Trial	Mean	SD	SE
Anthropomorphism	T1	2.21	0.69	0.11
	T2	2.17	0.69	0.11
	T3	2.35	0.86	0.13
Animacy	T1	2.74	0.74	0.12
	T2	2.87	0.85	0.13
	T3	2.78	0.79	0.12
Likeability	T1	3.57	0.70	0.11
	T2	3.40	0.52	0.08
	T3	3.50	0.73	0.11
Perceived intelligence	T1	3.49	0.79	0.12
	T2	3.50	0.71	0.11
	T3	3.50	0.79	0.12

Notes: SD = standard deviation; SE = standard error

Table G-19. Repeated measures ANOVA results for Godspeed Questionnaire Series scores

Measure	<i>F</i> (2,38)	Sig.	η^2_p
Anthropomorphism	1.21	0.298	0.03
Animacy	1.25	0.299	0.06
Likeability	2.97	0.063	0.14
Perceived intelligence	0.18	0.782	0.01

Notes: ANOVA = analysis of variance; Sig. = significance

Table G-20. Post hoc paired t-test results for Godspeed Questionnaire Series overall scores.

Measure	Trials	<i>t</i> (40)	Sig.	Cohen's <i>d</i>
Anthropomorphism	T1–T2	0.34	0.735	0.05
	T2–T3	–2.39	0.022	0.37
Animacy	T1–T2	–1.40	0.168	0.22
	T2–T3	1.38	0.174	0.22
Likeability	T1–T2	2.23	0.032	0.35
	T2–T3	–1.36	0.181	0.21
Perceived intelligence	T1–T2	–0.12	0.904	0.02
	T2–T3	–0.01	0.991	0.00

Note: Sig. = significance

Table G-21. GQS descriptive statistics, by learning condition (LC).

Condition	Measure	Trial	Mean	SD	SE
LC1	Anthropomorphism	T1	2.24	0.63	0.14
		T2	2.17	0.84	0.19
		T3	2.38	0.78	0.17
	Animacy	T1	2.79	0.55	0.12
		T2	2.91	0.84	0.19
		T3	2.82	0.72	0.16
	Likeability	T1	3.54	0.61	0.14
		T2	3.35	0.49	0.11
		T3	3.43	0.66	0.15
	Perceived intelligence	T1	3.42	0.64	0.14
		T2	3.37	0.73	0.16
		T3	3.37	0.67	0.15
LC2	Anthropomorphism	T1	2.17	0.74	0.16
		T2	2.41	0.97	0.21
		T3	2.32	0.95	0.21
	Animacy	T1	2.69	0.89	0.20
		T2	2.83	0.87	0.19
		T3	2.74	0.86	0.19
	Likeability	T1	3.59	0.79	0.17
		T2	3.43	0.52	0.11
		T3	3.56	0.81	0.18
	Perceived intelligence	T1	3.56	0.92	0.20
		T2	3.63	0.68	0.15
		T3	3.73	0.65	0.14

Notes: GQS = Godspeed Questionnaire Series;
SD = standard deviation; SE = standard error

Table G-22. Post hoc paired t-test results for GQS scores by learning condition (LC).

LC	Measure	Trials	<i>t</i> (<i>l</i>)	<i>df</i>	Sig.	Cohen's <i>d</i>
LC1	Anthropomorphism	T1–T2	0.60	19	0.557	0.09
		T2–T3	–1.74	19	0.099	0.26
	Animacy	T1–T2	–0.93	19	0.363	0.16
		T2–T3	0.79	19	0.440	0.12
	Likeability	T1–T2	1.34	19	0.197	0.35
		T2–T3	–0.27	19	0.788	0.15
	Perceived intelligence	T1–T2	0.29	19	0.772	0.07
		T2–T3	0.42	19	0.681	0.00
LC2	Anthropomorphism	T1–T2	1.02	20	0.320	0.27
		T2–T3	0.75	20	0.462	0.09
	Animacy	T1–T2	–1.37	20	0.185	0.15
		T2–T3	–0.76	20	0.458	0.10
	Likeability	T1–T2	0.23	20	0.817	0.24
		T2–T3	1.07	20	0.299	0.19
	Perceived intelligence	T1–T2	–0.80	20	0.432	0.10
		T2–T3	–2.10	20	0.048	0.14

Notes: GQS = Godspeed Questionnaire Series; Sig. = significance

Table G-23. GQS measure correlations with performance measures.

GQS measure	Performance measure	Trial	Pearson's r	Sig. (two-tailed)	95% CI (two-tailed) ^a	
					Lower	Upper
Anthropomorphism	Duration	T1	-0.25	0.110	-0.52	0.06
		T2	-0.03	0.867	-0.33	0.28
		T3	-0.04	0.782	-0.35	0.27
	Distance	T1	-0.26	0.095	-0.53	0.05
		T2	-0.01	0.962	-0.31	0.30
		T3	-0.06	0.707	-0.36	0.25
	Quality-MHD	T1	-0.15	0.343	-0.44	0.16
		T2	-0.11	0.510	-0.40	0.21
		T3	0.27	0.086	-0.04	0.53
	AutoQuality-MHD	T1	0.40	0.009	0.11	0.63
		T2	0.10	0.542	-0.22	0.39
		T3	-0.04	0.810	-0.34	0.27
	AutoDistance	T1	0.21	0.179	-0.10	0.49
		T2	-0.10	0.521	-0.40	0.21
		T3	-0.10	0.537	-0.39	0.21
Animacy	Duration	T1	-0.43	0.005	-0.65	-0.15
		T2	0.00	0.994	-0.31	0.31
		T3	-0.13	0.418	-0.42	0.19
	Distance	T1	-0.45	0.003	-0.67	-0.16
		T2	-0.02	0.899	-0.33	0.29
		T3	-0.14	0.392	-0.43	0.18
	Quality-MHD	T1	-0.13	0.402	-0.42	0.18
		T2	-0.07	0.646	-0.37	0.24
		T3	0.12	0.463	-0.20	0.41
	AutoQuality-MHD	T1	0.18	0.264	-0.14	0.46
		T2	0.08	0.632	-0.24	0.38
		T3	-0.10	0.536	-0.40	0.21
	AutoDistance	T1	0.37	0.017	0.07	0.61
		T2	-0.11	0.485	-0.41	0.20
		T3	-0.02	0.902	-0.33	0.29
Likeability	Duration	T1	-0.06	0.724	-0.36	0.26
		T2	0.16	0.331	-0.16	0.44
		T3	0.06	0.726	-0.26	0.36
	Distance	T1	-0.02	0.895	-0.33	0.29
		T2	0.10	0.551	-0.22	0.39
		T3	0.04	0.787	-0.27	0.35
	Quality-MHD	T1	-0.16	0.329	-0.44	0.16
		T2	-0.23	0.151	-0.50	0.09
		T3	0.03	0.864	-0.28	0.33
	AutoQuality-MHD	T1	0.15	0.340	-0.16	0.44
		T2	-0.08	0.613	-0.38	0.23
		T3	-0.09	0.560	-0.39	0.22
	AutoDistance	T1	0.01	0.926	-0.29	0.32
		T2	-0.28	0.077	-0.54	0.03
		T3	-0.24	0.129	-0.51	0.07

Notes: CI = confidence interval; GQS = Godspeed Questionnaire Series; MHD = modified Hausdorff distance; Sig. = significance

^a Estimation is based on Fisher's r-to-z transformation.

Table G-23. GQS measure correlations with performance measures (continued).

GQS measure	Performance measure	Trial	Pearson's r	Sig. (two-tailed)	95% CI (two-tailed) ^a	
					Lower	Upper
Perceived intelligence	Duration	T1	-0.09	0.570	-0.39	0.22
		T2	-0.11	0.503	-0.40	0.21
		T3	-0.07	0.653	-0.37	0.24
	Distance	T1	-0.12	0.460	-0.41	0.20
		T2	-0.15	0.360	-0.43	0.17
		T3	-0.05	0.753	-0.35	0.26
	Quality-MHD	T1	0.01	0.938	-0.30	0.32
		T2	-0.34	0.031	-0.58	-0.03
		T3	-0.11	0.503	-0.40	0.21
	AutoQuality-MHD	T1	0.42	0.007	0.12	0.64
		T2	-0.05	0.742	-0.35	0.26
		T3	-0.23	0.151	-0.50	0.09
	AutoDistance	T1	0.21	0.191	-0.11	0.48
		T2	-0.12	0.453	-0.41	0.19
		T3	0.11	0.488	-0.20	0.41

Notes: CI = confidence interval; GQS = Godspeed Questionnaire Series; MHD = modified Hausdorff distance; Sig. = significance

^a Estimation is based on Fisher's r-to-z transformation.

G-5 Robot Learning Questionnaire (RLQ) Results

Table G-24. RLQ scores, descriptive statistics.

Measure	Trial	Mean	SD	SE
Number of demonstrations	T1	3.84	0.37	0.06
	T2	3.58	0.66	0.10
	T3	3.58	0.73	0.11
Effectiveness	T1	2.81	0.42	0.06
	T2	2.93	0.48	0.07
	T3	3.01	0.62	0.09
Robot learning	T1	2.79	0.52	0.08
	T2	2.79	0.58	0.09
	T3	2.75	0.58	0.09
Improve	T1	2.90	0.79	0.12
	T2	2.79	0.81	0.13
	T3	2.88	0.80	0.12
Clarity	T1	2.67	0.65	0.10
	T2	2.69	0.72	0.11
	T3	2.86	0.68	0.11
Solo repeat	T1	2.67	0.69	0.11
	T2	2.86	0.72	0.11
	T3	2.95	0.58	0.09

Notes: SD = standard deviation; SE = standard error

Table G-25. Robot learning scores, repeated measure results

Measure	<i>F</i> (2,40)	Sig.	η^2_p
Demonstration	4.34	0.020	0.18
Effective	2.09	0.136	0.09
Robot learn	0.12	0.885	0.01
Improve	0.57	0.572	0.03
Clarity	1.29	0.287	0.06
Solo repeat	2.59	0.088	0.11

Note: Sig. = significance

Table G-26. Robot learning scores, paired t-test results.

Measure	Trials	<i>t</i> (1)	<i>df</i>	Sig.	Cohen's <i>d</i>
Demonstration	T1–T2	2.89	42	0.006	0.48
	T2–T3	0.00	42	1.000	0.00
Effective	T1–T2	–1.45	42	0.154	0.27
	T2–T3	–1.00	42	0.323	0.14
Robot learn	T1–T2	0.03	42	0.977	0.00
	T2–T3	0.45	42	0.659	0.06
Improve	T1–T2	1.00	42	0.323	0.15
	T2–T3	–0.85	41	0.400	0.12
Clarity	T1–T2	–0.20	42	0.844	0.03
	T2–T3	–1.36	41	0.181	0.24
Solo repeat	T1–T2	–1.19	42	0.241	0.23
	T2–T3	–0.85	41	0.400	0.15

Note: Sig. = significance

Table G-27. Robot learning scores descriptive statistics by learning condition (LC).

Measure	Trial	LC1 <i>N</i> = 21			LC2 <i>N</i> = 20		
		Mean	SD	SE	Mean	SD	SE
Number of demonstrations	T1	3.90	0.31	0.07	3.81	0.40	0.09
	T2	3.60	0.50	0.11	3.67	0.58	0.13
	T3	3.85	0.49	0.11	3.67	0.66	0.14
Effectiveness	T1	2.83	0.45	0.10	2.80	0.37	0.08
	T2	2.94	0.45	0.10	2.97	0.55	0.12
	T3	2.84	0.61	0.14	3.19	0.52	0.11
Robot learning	T1	2.69	0.49	0.11	2.91	0.47	0.10
	T2	2.64	0.53	0.12	2.94	0.58	0.13
	T3	2.59	0.59	0.13	2.96	0.46	0.10
Improve	T1	2.90	0.72	0.16	3.00	0.77	0.17
	T2	2.85	0.75	0.17	2.76	0.89	0.19
	T3	2.70	0.80	0.18	3.10	0.77	0.17
Clarity	T1	2.65	0.67	0.15	2.76	0.54	0.12
	T2	2.50	0.69	0.15	2.95	0.59	0.13
	T3	2.70	0.73	0.16	3.05	0.59	0.13
Solo repeat	T1	2.65	0.67	0.15	2.71	0.72	0.16
	T2	2.70	0.73	0.16	3.00	0.71	0.15
	T3	2.90	0.64	0.14	3.05	0.50	0.11

Notes: SD = standard deviation; SE = standard error

Table G-28. Robot learning scores, paired t-test results by learning condition (LC).

LC	Measure	Trials	<i>t</i> (<i>I</i>)	<i>df</i>	Sig.	Cohen's <i>d</i>
LC1	Demonstration	T1–T2	2.85	19	0.010	0.72
		T2–T3	–2.03	19	0.056	0.50
	Effective	T1–T2	–0.94	19	0.360	0.25
		T2–T3	0.85	19	0.406	0.20
	Robot learn	T1–T2	0.48	19	0.638	0.10
		T2–T3	0.38	19	0.707	0.09
	Improve	T1–T2	0.29	19	0.772	0.07
		T2–T3	1.14	19	0.267	0.19
	Clarity	T1–T2	0.77	19	0.453	0.22
		T2–T3	–1.16	19	0.258	0.28
	Solo repeat	T1–T2	–0.29	19	0.772	0.07
		T2–T3	–1.16	19	0.258	0.29
LC2	Demonstration	T1–T2	1.37	20	0.186	0.29
		T2–T3	0.00	20	1.000	0.00
	Effective	T1–T2	–1.35	20	0.192	0.37
		T2–T3	–2.53	20	0.020	0.41
	Robot learn	T1–T2	–0.22	20	0.827	0.05
		T2–T3	–0.12	20	0.905	0.03
	Improve	T1–T2	1.42	20	0.171	0.29
		T2–T3	–1.92	20	0.069	0.40
	Clarity	T1–T2	–1.28	20	0.214	0.34
		T2–T3	–0.53	20	0.605	0.16
	Solo repeat	T1–T2	–1.30	20	0.208	0.40
		T2–T3	–0.33	20	0.748	0.08

Note: Sig. = significance

Table G-29. RLQ factor correlations with performance measures.

RLQ factor	Performance measure	Trial	Pearson's <i>r</i>	Sig. (two-tailed)	95% confidence interval (two-tailed) ^a	
					Lower	Upper
Number of demonstrations	Duration	T1	–0.36	0.017	–0.60	–0.07
		T2	–0.07	0.644	–0.36	0.23
		T3	–0.18	0.257	–0.45	0.13
	Distance	T1	–0.36	0.018	–0.60	–0.07
		T2	–0.10	0.530	–0.39	0.21
		T3	–0.12	0.446	–0.41	0.19
	Quality-MHD	T1	–0.32	0.039	–0.56	–0.02
		T2	–0.15	0.331	–0.43	0.16
		T3	0.30	0.048	0.00	0.55
	AutoQuality-MHD	T1	–0.30	0.051	–0.55	0.00
		T2	–0.22	0.154	–0.49	0.08
		T3	–0.34	0.026	–0.58	–0.04
	AutoDistance	T1	0.29	0.058	–0.01	0.54
		T2	–0.49	0.001	–0.69	–0.23
		T3	–0.42	0.005	–0.64	–0.14

Notes: MHD = modified Hausdorff distance; RLQ = Robot Learning Questionnaire; Sig. = significance

^a Estimation is based on Fisher's *r*-to-*z* transformation.

Table G-29. RLQ factor correlations with performance measures (continued).

RLQ factor	Performance measure	Trial	Pearson's r	Sig. (two-tailed)	95% confidence interval (two-tailed) ^a	
					Lower	Upper
How effective was your last demonstration? Effectiveness	Duration	T1	0.21	0.176	-0.10	0.48
		T2	-0.29	0.063	-0.54	0.02
		T3	-0.04	0.809	-0.33	0.27
	Distance	T1	0.17	0.283	-0.14	0.45
		T2	-0.24	0.122	-0.50	0.07
		T3	0.02	0.900	-0.28	0.32
	Quality-MHD	T1	-0.22	0.155	-0.49	0.09
		T2	0.35	0.021	0.06	0.59
		T3	0.20	0.191	-0.10	0.47
	AutoQuality-MHD	T1	-0.09	0.558	-0.38	0.21
		T2	-0.02	0.906	-0.32	0.28
		T3	-0.13	0.416	-0.41	0.18
	AutoDistance	T1	-0.03	0.825	-0.33	0.27
		T2	0.32	0.036	0.02	0.57
		T3	-0.01	0.952	-0.31	0.29
How well has the robot learned the desired behavior? Robot learning	Duration	T1	0.32	0.039	0.02	0.56
		T2	-0.10	0.518	-0.39	0.21
		T3	0.12	0.445	-0.19	0.41
	Distance	T1	0.26	0.097	-0.05	0.52
		T2	-0.06	0.706	-0.35	0.25
		T3	0.12	0.456	-0.19	0.40
	Quality-MHD	T1	-0.30	0.050	-0.55	0.00
		T2	0.18	0.253	-0.13	0.45
		T3	-0.07	0.675	-0.36	0.24
	AutoQuality-MHD	T1	0.04	0.800	-0.26	0.34
		T2	-0.13	0.417	-0.41	0.18
		T3	-0.22	0.152	-0.49	0.08
	AutoDistance	T1	0.02	0.913	-0.28	0.32
		T2	0.44	0.003	0.16	0.65
		T3	0.00	0.985	-0.30	0.30
Would additional demonstrations continue to improve robot behavior?	Duration	T1	0.02	0.880	-0.28	0.32
		T2	0.00	0.990	-0.30	0.30
		T3	-0.07	0.638	-0.37	0.23
	Distance	T1	-0.03	0.863	-0.32	0.28
		T2	-0.07	0.639	-0.37	0.23
		T3	-0.08	0.608	-0.38	0.23
	Quality-MHD	T1	-0.01	0.949	-0.31	0.29
		T2	-0.05	0.737	-0.35	0.25
		T3	-0.16	0.303	-0.44	0.15
	AutoQuality-MHD	T1	-0.27	0.079	-0.53	0.03
		T2	-0.04	0.821	-0.33	0.27
		T3	-0.09	0.551	-0.39	0.22
	AutoDistance	T1	-0.08	0.609	-0.37	0.23
		T2	0.20	0.206	-0.11	0.47
		T3	0.08	0.596	-0.23	0.38

Notes: MHD = modified Hausdorff distance; RLQ = Robot Learning Questionnaire; Sig. = significance

^a Estimation is based on Fisher's r-to-z transformation.

Table G-29. RLQ factor correlations with performance measures (continued).

RLQ factor	Performance measure	Trial	Pearson's r	Sig. (two-tailed)	95% confidence interval (two-tailed) ^a	
					Lower	Upper
How clear was the robot's understanding of the demonstration intent? Clarity	Duration	T1	0.18	0.237	-0.12	0.46
		T2	-0.08	0.622	-0.37	0.23
		T3	-0.21	0.185	-0.48	0.10
	Distance	T1	0.08	0.628	-0.23	0.37
		T2	-0.10	0.543	-0.38	0.21
		T3	-0.23	0.142	-0.50	0.08
	Quality-MHD	T1	-0.07	0.653	-0.36	0.23
		T2	0.18	0.238	-0.12	0.46
		T3	0.00	0.977	-0.31	0.30
	AutoQuality-MHD	T1	-0.05	0.762	-0.34	0.26
		T2	-0.19	0.226	-0.46	0.12
		T3	-0.21	0.186	-0.48	0.10
	AutoDistance	T1	0.00	0.997	-0.30	0.30
		T2	0.20	0.197	-0.11	0.47
		T3	0.32	0.041	0.01	0.57
How well would the robot traverse a similar new path unassisted? Solo repeat	Duration	T1	0.09	0.564	-0.22	0.38
		T2	-0.15	0.331	-0.43	0.16
		T3	-0.23	0.138	-0.50	0.08
	Distance	T1	-0.01	0.950	-0.31	0.29
		T2	-0.20	0.198	-0.47	0.11
		T3	-0.25	0.112	-0.51	0.06
	Quality-MHD	T1	-0.03	0.830	-0.33	0.27
		T2	0.02	0.877	-0.28	0.32
		T3	-0.17	0.277	-0.45	0.14
	AutoQuality-MHD	T1	0.09	0.583	-0.22	0.38
		T2	-0.47	0.002	-0.67	-0.19
		T3	-0.01	0.969	-0.31	0.30
	AutoDistance	T1	0.16	0.294	-0.14	0.44
		T2	0.35	0.022	0.05	0.59
		T3	0.44	0.003	0.16	0.66

Notes: MHD = modified Hausdorff distance; RLQ = Robot Learning Questionnaire; Sig. = significance

^a Estimation is based on Fisher's r-to-z transformation.

List of Symbols, Abbreviations, and Acronyms

AI	Artificial Intelligence
ANOVA	analysis of variance
DEVCOM	U.S. Army Combat Capabilities Development Command
GQS	Godspeed Questionnaire Series
LC	learning condition
LiDAR	Light Detection and Ranging
MHD	modified Hausdorff distance
ML	machine learning
NASA	National Aeronautics and Space Administration
PEBL	Psychology Experiment Building Language
RL	reinforcement learning
RLQ	Robot Learning Questionnaire
SD	standard deviation
SE	standard error
T	transparency
T1	transparency: baseline
T2	transparent intent
T3	transparent impact
TIA	Trust in Automation
TLX	Task Load Index

1 DEFENSE TECHNICAL
(PDF) INFORMATION CTR
DTIC OCA

1 DEVCOM ARL
(PDF) FCDD RLB CI
TECH LIB