Evaluating the effects of enhanced autonomy transparency on trust, dependence, and human-autonomy team performance over time

Ruikun Luo

Robotics Institute, University of Michigan, Ann Arbor

Na Du

Department of Informatics and Networked Systems, University of Pittsburgh

X. Jessie Yang

Industrial and Operations Engineering, University of Michigan, Ann Arbor

Manuscript type: *Original Research*

Running head: Enhancing autonomy transparency

Word count: 4500

Corresponding author: X. Jessie Yang, 1205 Beal Avenue, Ann Arbor, MI48109, Email: xijyang@umich.edu

Acknowledgement: This work is partially supported by the National Science Foundation under Grant No. 2045009. We would also like to thank Kevin Y. Huang for his assistance in data collection.

Abstract

As autonomous systems become more complicated, humans may have difficulty deciphering autonomy-generated solutions and increasingly perceive autonomy as a mysterious black box. The lack of transparency contributes to the lack of trust in autonomy and suboptimal team performance. In response to this concern, researchers have proposed various methods to enhance autonomy transparency and evaluated how enhanced transparency could affect the people's trust and the human-autonomy team performance. However, the majority of prior studies measured trust at the end of the experiment and averaged behavioral and performance measures across all trials in an experiment, yet overlooked the temporal dynamics of those variables. We have little understanding of how autonomy transparency affects trust, dependence, and performance *over time*. The present study, therefore, aims to fill the gap and examine such temporal dynamics. We develop a game *Treasure Hunter* wherein a human uncovers a map for treasures with the help from an intelligent assistant. The intelligent assistant recommends where the human should go next. The rationale behind each recommendation could be conveyed in a display that explicitly lists the option space (i.e., all the possible actions) and the reason why a particular action is the most appropriate in a given context. Results from a human-in-the-loop experiment with 28 participants indicate that by conveying the intelligent assistant's decision-making rationale via the display, participants' trust increases significantly and become more calibrated over time. Using the display also leads to a higher acceptance of recommendations from the intelligent agent.

Keyword: Transparent autonomy, Human-autonomy interaction, Human-automation interaction, Design rationale, Trust calibration, Propositional logic.

1. INTRODUCTION

While the advances in artificial intelligence and machine learning empower a new generation of autonomous systems for assisting human performance, one major concern arises from the human factors perspective: Human agents have difficulty deciphering autonomy-generated solutions and increasingly perceive autonomy as a mysterious black box. The lack of transparency contributes to the lack of trust in autonomy and suboptimal team performance (Chen & Barnes, 2014; de Visser, Pak, & Shaw, 2018; Du et al., 2019; Endsley, 2017; Lyons & Havig, 2014; Lyons et al., 2016; Yang, Unhelkar, Li, & Shah, 2017).

In response to this concern, researchers have investigated ways to enhance autonomy transparency. Research has shown that conveying the system's reliability, confidence, performance, and reason for actions and errors, even in hand-crafted forms, can facilitate the establishment of trust and improve human-autonomy team performance (Dzindolet, Peterson, Pomranky, Pierce, & Beck, 2003; Koo et al., 2014; Koo, Shin, Steinert, & Leifer, 2016; Mercado et al., 2016; Seong & Bisantz, 2008; Wang, Jamieson, & Hollands, 2009; Yang et al., 2017).

These studies provide valuable insights in the design of transparent autonomy. However, the majority of prior literature evaluated trust through questionnaires administered at the end of an experiment, averaged behavioral and performance measures over all trials, yet overlooked the temporal dynamics of these variables when people interact with autonomy repeatedly. Trust is a dynamic variable that change as a result of moment-to-moment interaction with the autonomous system (Guo & Yang, 2021; Yang, Schemanske, & Searle, 2021), and so are people's dependence behaviors and the human-autonomy team performance. This study, therefore, aims to examine how autonomy transparency affects people's trust and dependence behavior, and human-autonomy team performance *over time.*

2. RELATED WORK

In this section, we review prior literature examining the impact of various methods aimed at enhancing autonomy transparency on trust, dependence behaviors, and human-autonomy team performance. It is worth to note that in almost all the studies, human participants interacted with automated/autonomous systems over multiple trials. However, trust was usually measured once at the end of the experiment (i.e., the "snapshot of trust (Guo & Yang, 2021)"). Behavioral and performance measures were usually averaged over all the trials.

There are multiple definitions of autonomy transparency, to name a few: "the [degree of] shared intent and shared awareness between a human and a machine (Lyons & Havig, 2014)", "the extent to which an autonomous agent can convey its intent, performance, future plans and reasoning process (Chen et al., 2014)", "a mechanism to expose the decision-making of a robot (Theodorou, Wortham, & Bryson, 2017)", "the understandability and predictability of their actions (Endsley, 2017)", "the ability for the automation to be inspectable or viewable in the sense that its mechanisms and rationale can be readily known (Miller, 2018)". Despite the lack of a universal definition, a consistent pattern can be observed: a transparent autonomy should communicate to the human agent the autonomy's ability and performance, its decision-making logic and rationale, and its intent and future plans.

Although autonomy transparency was only recently defined, research has been conducted to convey certain aspects of autonomy-generated solutions. One body of human factors research has concentrated on conveying likelihood information in the form of automation reliability, (un)certainty, and confidence. Some studies revealed that likelihood information significantly helped people calibrate their trust and enhance human-automation team performance (McGuirl & Sarter, 2006; Walliser, de Visser, & Shaw, 2016; Wang et al., 2009). McGuirl and Sarter (2006) examined whether providing the updated system confidence information of an automated decision support system (DSS) improves human-autonomy team performance. In their experiment, 15 instructor pilots flew a series of 28 short flights. The DSS assisted pilots with detecting and handling in-flight icing encounters. Results showed that pilots experienced significantly fewer icing-related stalls with the updated confidence information. In the study of Wang et al. (2009), participants performed a target detection task, identifying friends from foes with the aid from an imperfect combat identification (CID) system. Results showed that when the reliability of the CID system was disclosed to the participants, they had better reliance behaviors, indicated by a more optimal response bias β. In contrast to the supportive evidence, other studies showed that presenting likelihood information did not lead to more appropriate trust, nor better performance (Bagheri & Jamieson, 2004; Fletcher, Bartlett, Cockshell, & McCarley, 2017). For example, Bagheri and Jamieson (2004) examined the impact of providing operators with information about automation reliability. In their experiment, participants performed three tasks simultaneously, one of which was assisted by an automated decision aid. However, sometimes the automated decision aid would fail (miss) to complete the task and the human should intervene. Automation reliability, essentially the hit rate ("Slightly under 100%" or "Slightly above 50%") was disclosed to the participants. Contrasting the results from the study to their previous study where participants were unaware of the hit rate suggested no benefits of disclosing hit rate on trust and task performance. To explain the seemingly mixed results, Du, Huang, and Yang (2020) summarized three different types of likelihood information that has been tested in prior literature: positive and negative predictive values, overall success likelihood, and hit and correct rejection rates and hypothesized that not all likelihood information is equal in aiding human-autonomy team performance. They conducted a human-subject experiment with 60 participants using a simulated surveillance task. Each participant performed a compensatory tracking task and a threat detection task with the help of an imperfect automated threat detector. The three types of likelihood information were presented. Results showed that presenting the predictive values or the overall likelihood value, rather than the hit and correct rejection rates, leads to more appropriate reliance behaviors and higher human-autonomy task performance.

Besides conveying likelihood information, the second body of research has investigated the impact of providing hand-crafted explanations of autonomy's behaviors. For example, Dzindolet et al. (2003) conducted a study where participants detected the presence or absence of a camouflaged soldier when viewing slides of Fort Sill terrain, with the help of an imperfect contrast detector. Results of their study showed that after observing an automation failure, participants' trust decreased unless they were provided with a hand-crafted explanations on why a decision aid might err ("The contrast detector will indicate the soldier is present if it detects forms that humans often take. Since non-humans (e.g., shading from a tree) sometimes take human-like forms, mistakes can be made."). In the context of autonomous driving, the studies of Koo et al. (2014) and Koo et al. (2016) showed that informing the drivers of the hand-crafted reasons for automated braking (e.g., road hazard ahead) decreased drivers' anxiety and increased their sense of control, preference, and acceptance. Similarly, Du et al. (2019) found that speech output explaining why and how the automated vehicle is going to take certain actions was rated higher on trust, preference, usability and acceptance.

More recently, research has formally defined autonomy transparency. Notably, Mercado et al. (2016) and Chen et al. (2018) proposed the situation awareness-based agent transparency (SAT) model to convey information supporting the human agent's perception, comprehension, and projection of an intelligent assistant's recommendations. SAT level 1 conveys the agent's current status/action/plans; SAT level 2 describes the agent's reasoning process; SAT level 3 presents the agent's projections/predictions. In the study of Mercado et al. (2016), participants controlled a group of heterogeneous unmanned vehicles (UxVs). An intelligent agent assisted the participant and provided suggestions. Three transparency levels were introduced: Level 1 was the baseline condition showing basic plan information. Level 1+2 contained all the information provided in Level 1 plus the agent's reasoning and rationale behind recommending the plans. Level 1+2+3 contained all the information provided in Levels 1 and 2 plus projection of uncertainty information. Results showed that as the transparency level increased, participants' trust in and perceived usability of the intelligent agent increased significantly, and so did the human-agent team performance.

The above-mentioned studies provide valuable insights on whether and how transparent autonomy could enhance people's trust and dependence, and the human-autonomy team performance. However, the majority of prior studies measured trust at a snapshot, usually at the end of the experiment. More recently, researchers started to emphasize the importance of viewing trust as a dynamic variable and examining the temporal dynamics when a human interacts with an autonomous system over time (de Visser et al., 2020; Guo & Yang, 2021; Yang et al., 2021, 2017). Yet, we have little understanding on how autonomy transparency affect trust, dependence, and human-autonomy team performance over time. The present study, therefore, aims to fill the gap and examine the temporal dynamics.

3. METHOD

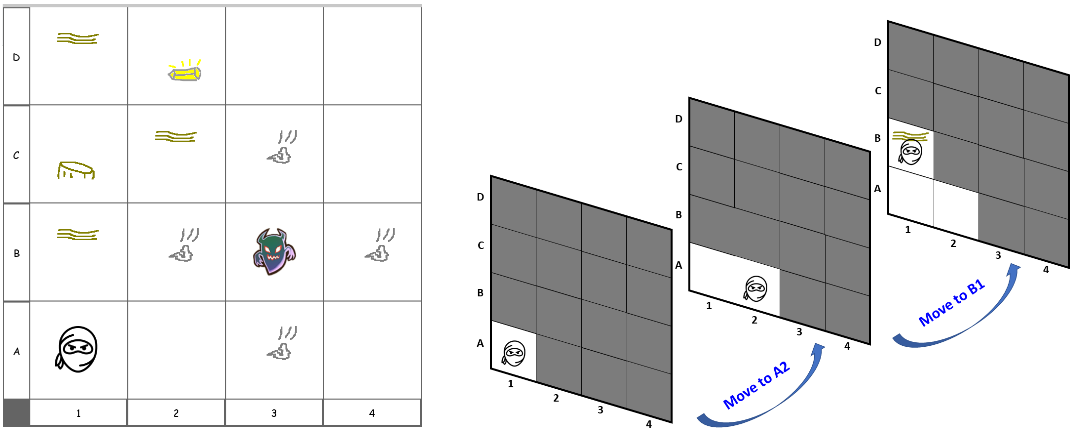
This research complied with the American Psychological Association code of ethics and was approved by the Institutional Review Board at the University of Michigan.

3.1 Participants

Thirty-four participants (Age: Mean = 21.17 years, *SD* = 1.66 years) took part in the experiment. All participants had normal or corrected-to-normal sight and hearing. Participants were compensated with $5 upon completion of the experiment. In addition, there was a chance to obtain an additional bonus of 1 to 20 dollars based on their performance.

3.2 Simulation testbed

We developed an experimental testbed – *Treasure Hunter*, adapted from the Wumpus world game (Russell & Norvig, 2010). In the game, the participant acts as a hunter to find the gold bar in the map with the help of an intelligent assistant (Figures 1a & 1b). Each step, the hunter can move to an un-visited location that is connected to the visited locations. Figure 1b shows that the hunter moves from A1 to A2 and then to B1. On the way to the treasure, the hunter might fall into a pit (shown in C1 in Figure 1a) or encounter a wumpus (shown in B3 in Figure 1a). The hunter gathers information about his or her surroundings by a set of sensors. The sensors will report a stench when the wumpus is in an adjacent location (shown as B2, A3, C3, B4 in Figure 1a) and a breeze when a pit is in an adjacent location (shown as B1, C2, D1 in Figure 1a). There is one and only one gold bar/wumpus on a map. However, there might be one or multiple pits in a map. Each element - a pit, a wumpus, or a gold bar - occupies a unique location on the map.



(a) (b)

Figure 1. (a) An example map in Treasure Hunter. Each square is denoted by the row number (from 1 to 4) and the column number (from A to D). (b) First two steps of a hunter moving in the map.

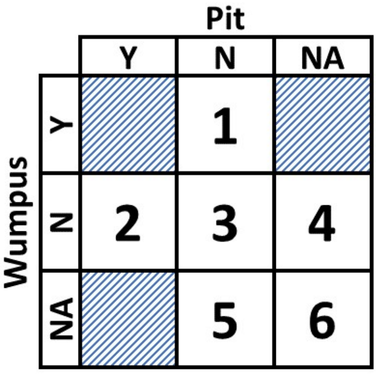
Table 1 shows the scores and consequences for different events. If the hunter finds the gold bar, s/he will receive 500 points, and the game will end. If the hunter encounters the wumpus, s/he will lose 1000 points, and the game will end. If the hunter falls into a pit, s/he will lose 100 points but can still continue the game. The hunter will only fall into a pit at the first time s/he encounters it. The hunter will get a 10-point penalty for uncovering every new location.

TABLE 1: *Scores and consequences for different events*

Event	Score	Consequence
Find the gold bar	+500	Map ends
Discover one new location	-10	Continue
Fall into a pit	-100	Continue, no more points lost when revisit
Meet wumpus	-1000	Map ends

An intelligent assistant helps the participant by recommending where to go. The intelligent assistant is a knowledge-based agent and reasons using propositional logic (Russell & Norvig, 2010). Propositional logic is a mathematical model that reasons about the truth or falsehood of logical statements. By using logical inference, the agent will give the values of four logical statements for a given location (e.g. location $D2$): (1) there is a pit at this location, denoted as $P_{D,2}$; (2) there is no pit at this location, denoted as $\neg P_{D,2}$; (3) there is a wumpus at this location, denoted as $W_{D,2}$; (4) there is no wumpus at this location, denoted as $\neg W_{D,2}$. Based on the value of these 4 logical statements, we can categorize the location into one of the six different conditions shown in Figure 2: Y represents there is a pit/wumpus at this location (value of the first/third logical statements is true); N represents there is no pit/wumpus at this location (value of the second/fourth logical statement is true); NA represents the agent is not sure about the existence of pit/wumpus at this location (values of all the four statements are false). The shaded squares in Figure 2 are the impossible cases because the pit and wumpus cannot co-exist in one location. For each case in Figure 2, the agent will assign probabilities of encountering a wumpus, falling into a pit, finding a gold bar or nothing happens as well as the corresponding expected scores if the hunter moves to that location as shown in Table 2. The agent will randomly select one of the potential next locations with the highest expected score as the recommendation.

Every step during the experiment, the participant will first receive the suggestion from the intelligent assistant, and then make a decision, i.e., select the target location that s/he wants to go next. After the participant makes a decision, the hunter will move

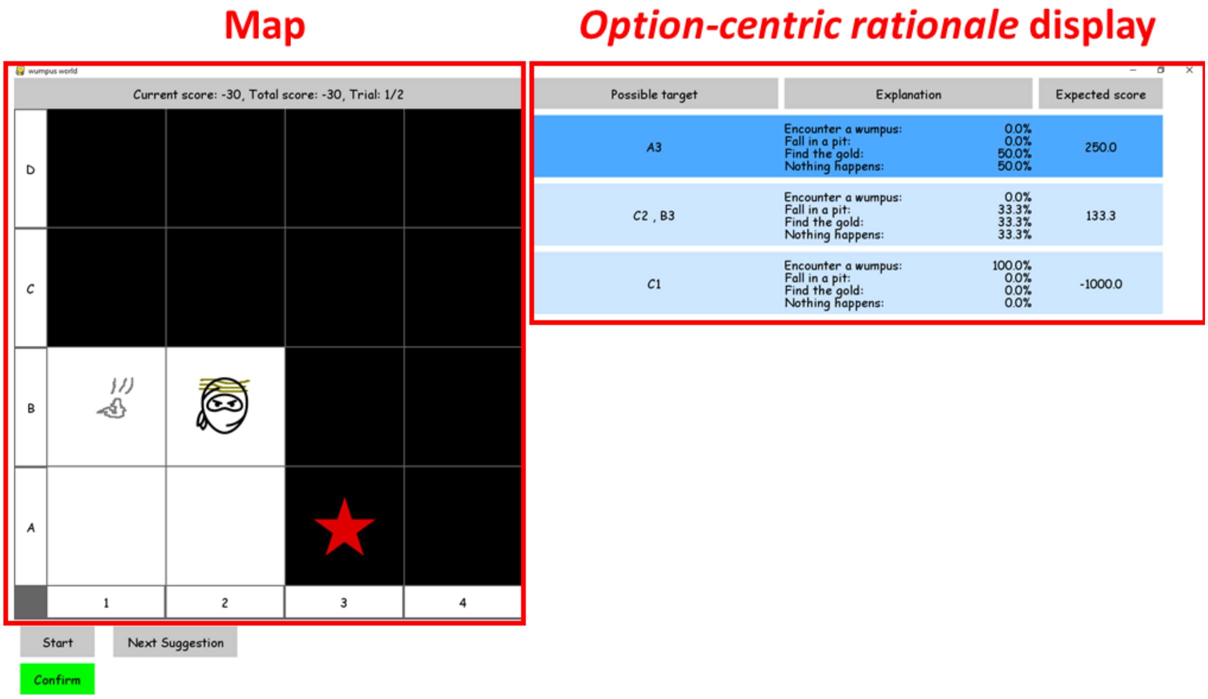


ID	P(W, P, G, N)	Expected Score
1	$(1, 0, 0, 0)$	-1000
2	$(0, 1, 0, 0)$	-100
3	$(0, 0, 0.5, 0.5)$	250
4	$(0, \frac{1}{3}, \frac{1}{3}, \frac{1}{3})$	133.33
5	$(\frac{1}{3}, 0, \frac{1}{3}, \frac{1}{3})$	-166.67
6	$(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$	-150

Figure 2 & Table 2. Six potential cases with the corresponding probabilities/expected scores based on the reasoning of pit and wumpus conditions from the intelligent assistant. Y: there is a pit/wumpus. N: there is no pit/wumpus. NA: it is not sure to have a pit/wumpus. Shaded squares are the impossible cases because the pit and wumpus cannot co-exist in one location. W: encounter a wumpus; P: fall into a pit; G: find a gold bar; N: nothing happens.

to the next location and the intelligent assistant will update its knowledge based on the sensory feedback (breeze and stench at the new location). The intelligent agent is a level 3/4 automation as defined by Sheridan and Verplank (1978) where the autonomy narrows down the action selections and suggests one alternative to the human agent.

In this study, we propose the option-centric rationale display for enhancing autonomy transparency. Inspired by the research on design rationale, the option-centric rationale display explicitly displays the option space (i.e., all the possible options/actions that an autonomy could take) and the rationale why a particular option is the most appropriate at a given context. Design rationale is an area of design science focusing on the "representation for explicitly documenting the reasoning and argumentation that make sense of a specific artifact (MacLean, Young, Bellotti, & Moran, 1991)". Its primary goal is to support designers and other stakeholders by recording the argumentation and reasoning behind the design process. The theoretical underpinning for design rationale is that for designers what is important is not just the specific artifact itself but its other possibilities – why an artifact is designed in a particular way compared to how it might otherwise be.



Current score: -30, Total score: -30, Trial: 1/2

Possible target	Explanation		Expected score
A3	Encounter a wumpus: Fall in a pit: Find the gold: Nothing happens:	0.0% 0.0% 50.0% 50.0%	250.0
C2 , B3	Encounter a wumpus: Fall in a pit: Find the gold: Nothing happens:	0.0% 33.3% 33.3% 33.3%	133.3
C1	Encounter a wumpus: Fall in a pit: Find the gold: Nothing happens:	100.0% 0.0% 0.0% 0.0%	-1000.0

Figure 3. Testbed with the option-centric rationale display.

Figure 3 shows the option-centric rationale display proposed in this study. The display details all the available next locations and the criteria for choosing a particular location, and highlights the final recommendation using a red star. The criteria for recommending a particular location depends on whether the human-autonomy team will find the gold bar, fall into a pit, encounter a wumpus, or uncover a new location (without finding a gold bar, falling into a pit or encountering a wumpus). The display also shows the possibility of each criterion and the corresponding expected score. The display will group the next locations based on the criteria, i.e., if two locations have the same probabilities of each criterion, the display will list them in the same row. The locations are sorted from the highest expected score to the lowest. The final recommendation is one of the locations with the highest expected score. Note that the available next locations, the possibility of each criterion and the expected scores are all computed by the intelligent assistant.

3.3 Experimental design

The experiment used a within-subjects design. The independent variables in the experiment were the presence/absence of the option-centric rationale display and time (i.e., trial number). The order of the presence/absence of the display was counterbalanced to eliminate potential order effects. Participants played the game on 5 different maps each, with and without the display. When the display was absent, the participant only saw a red star that indicated the recommendation by the intelligent assistant.

3.4 Measures

We measured three types of dependent variables: subjective responses, behavioral responses and performance. After completing each map, participants were asked to report their trust in the intelligent assistant and their self-confidence to accomplish the task without the intelligent assistant using two 9-point Likert scales: (1) How much do you trust the intelligent assistant? (2) How confident are you in completing tasks without the intelligent assistant?

We calculated the recommendation acceptance as the rate that the participant followed the recommendations given by the intelligent assistant. Participants' scores for each map were recorded as well.

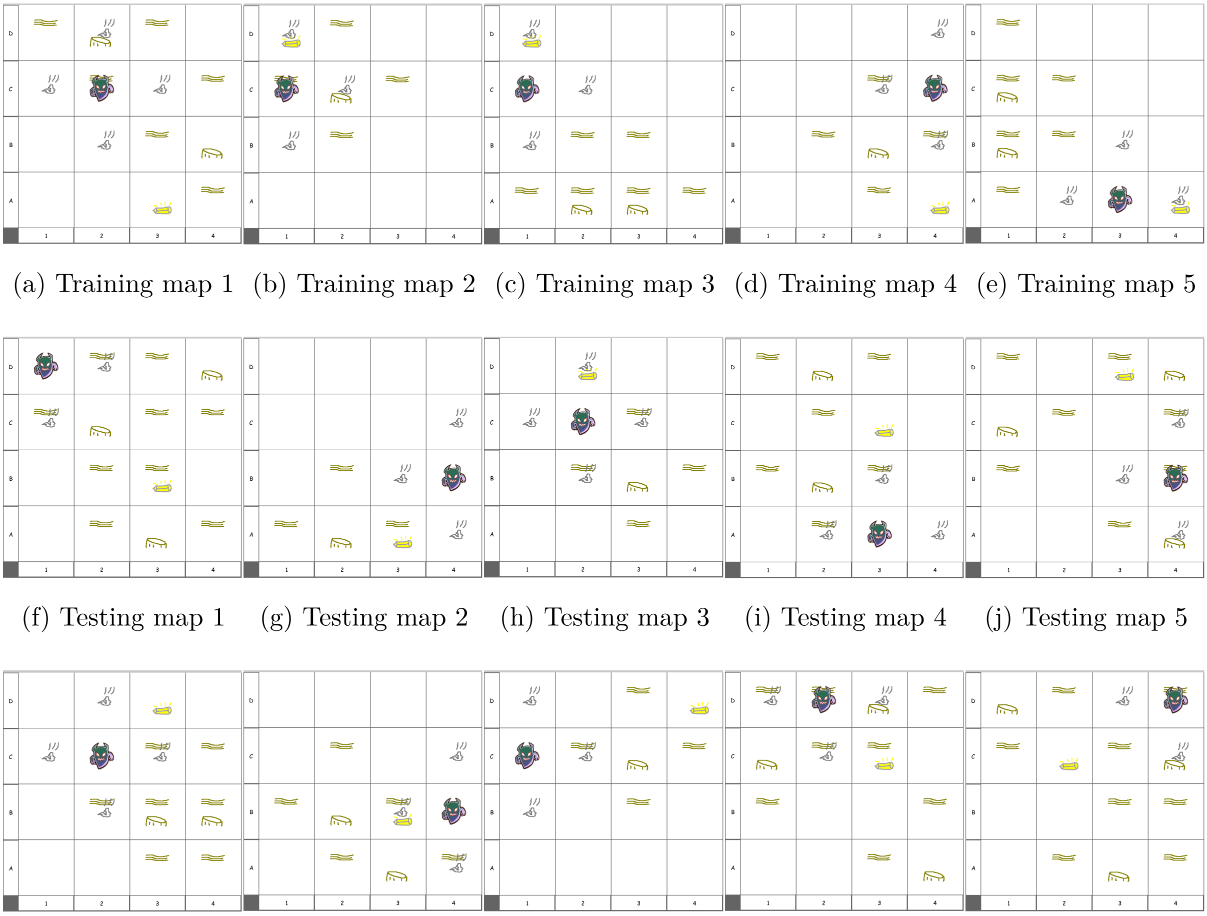
3.5 Map selection

In order to eliminate the inherent randomness of the task, we carefully selected the maps used in the experiment (Figure 4). First, we randomly generated 100 maps and ran the game only with the intelligent assistant 20 times for each map (i.e., always accepted the recommendations from the intelligent assistant). We ranked the maps based on the standard deviation of the scores for each map from the lowest to the highest. Second, we selected 10 maps that satisfied three criteria: (1) Each map had a low standard deviation of the scores; (2) In each map, the gold bar was not just next to the start location; (3) The locations of the gold bar in the 10 maps should be balanced across the maps instead of concentrating in one part of the maps (e.g. upper right corner of the map). For each participant, the order of the 10 maps in the experiment are randomly determined. The second row in Table 3 shows the mean and standard error of the intelligent assistant's score of the 10 selected maps.

We also developed 5 maps for the training session. Out of the 5 training maps, there are two maps with a pit next to the start location and three maps with low standard deviation of scores. The 5 training maps were presented according to the following order: The first was similar to the maps participants experience in the real test. Participants practiced on this map without the help of the intelligent assistant. The aim was to help participants get familiar with the game. From the second map onward, participants played the game with the help of the intelligent assistant. The second and fourth practice maps were similar to the maps participants experienced in the real test. The third and the fifth maps contained a pit next to the start location. The reason for selecting the two maps (i.e., the third and the fifth map) was to help participants fully understand the stochasticity of the game. For example, in the fifth training map (Figure 4), a breeze was detected by the sensor at the start location and the two adjacent locations (i.e., B1 and A2) have the same probability of having a pit.

3.6 Procedure

All participants provided informed consent and filled in a demographics survey. After that, participants received a practice session. Participants played the game first without the intelligent assistant, and practiced on another four maps with the intelligent assistant, and with or without the option-centric rationale display. In the experiment, participants played the game with 5 maps in each condition. Participants were told that the intelligent assistant in each condition was different and independent. After each map, participants were asked to report their trust in the intelligent assistant and their confidence in accomplishing the game without the help of the intelligent assistant. Participants' acceptance behaviors and task performance were recorded automatically by the testbed.



(a) Training map 1 (b) Training map 2 (c) Training map 3 (d) Training map 4 (e) Training map 5

(f) Testing map 1 (g) Testing map 2 (h) Testing map 3 (i) Testing map 4 (j) Testing map 5

(k) Testing map 6 (l) Testing map 7 (m) Testing map 8 (n) Testing map 9 (o) Testing map 10

Figure 4. Selected maps for training and testing. First row: fixed order training map. Second and third row: testing map, order is randomly determined for each participant.

4. RESULTS

Data from 4 participants were discarded due to malfunction of the testbed. Data from 2 participants were discarded as their task performance were considered as outliers based on the two-sided Dixon's Q test (Dixon, 1953). All analyses were conducted using data from the remaining 28 participants (Mean age = 21.25 years, SD = 1.72 years).

To examine how trust, dependence and performance changes over time with or without the option-centric rationale display, we conducted analyses using the mixed linear model. The model is particularly useful with repeated measurements from the same subject. The presence/absence of the display and trial number are modeled as fixed effects and intercepts for subjects as random effects. Results are reported as significant for $\alpha < .05$.

4.1 Trust over time with and without the option-centric rationale display

There was a significant interaction effect between the presence/absence of the option-centric rational display and time, $t(1, 249) = 2.38, p = .018$ (Figure 5). Specifically, when participants were provided with the display, their trust in the autonomous agent increased as they gained more experiences, $t(1, 111) = 6.59, p < .001$. When the display was absent, however, their trust did not change significantly, $t(1, 111) = 0.73, p = 0.46$.

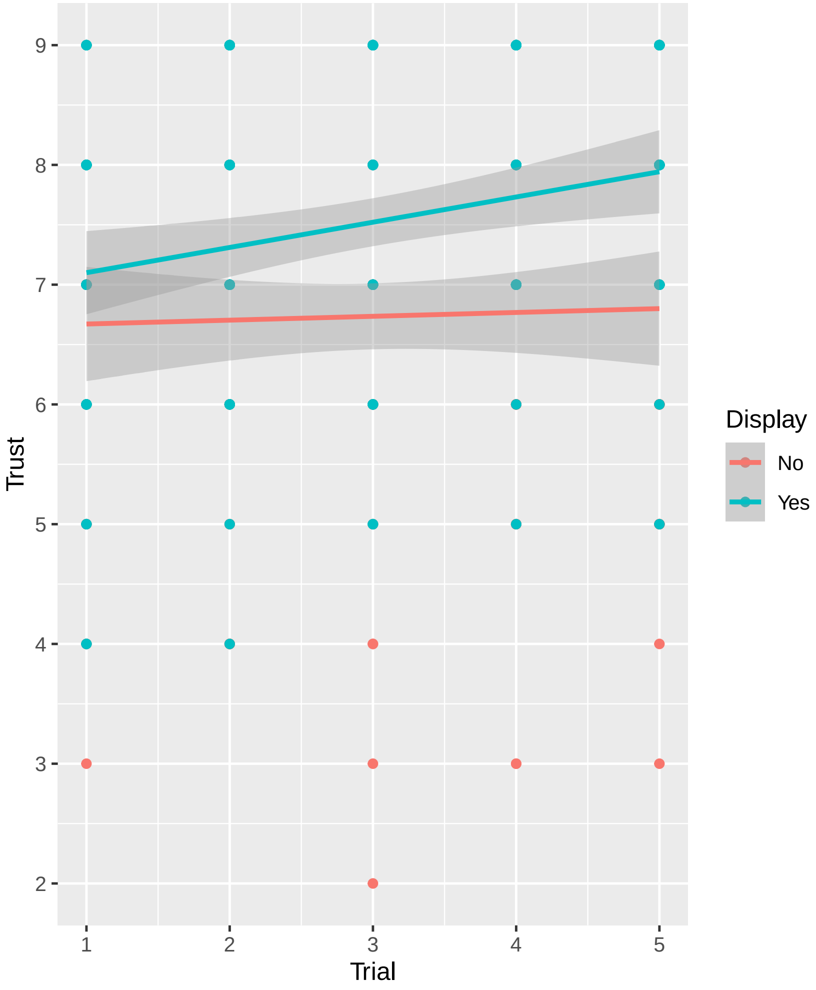


Figure 5. Trust in the intelligent agent over time with and without the option-centric rationale display.

4.1 Recommendation acceptance over time with and without the option-centric rationale display

There was a significant effect of display on the participant's acceptance rate, $t(1, 250) = 2.59, p = .01$ (Figure 6). When participants were provided with the display, there was a higher acceptance rate of the recommendations. The effect of time was not significant, $t(1, 250) = -0.30, p = .76$.

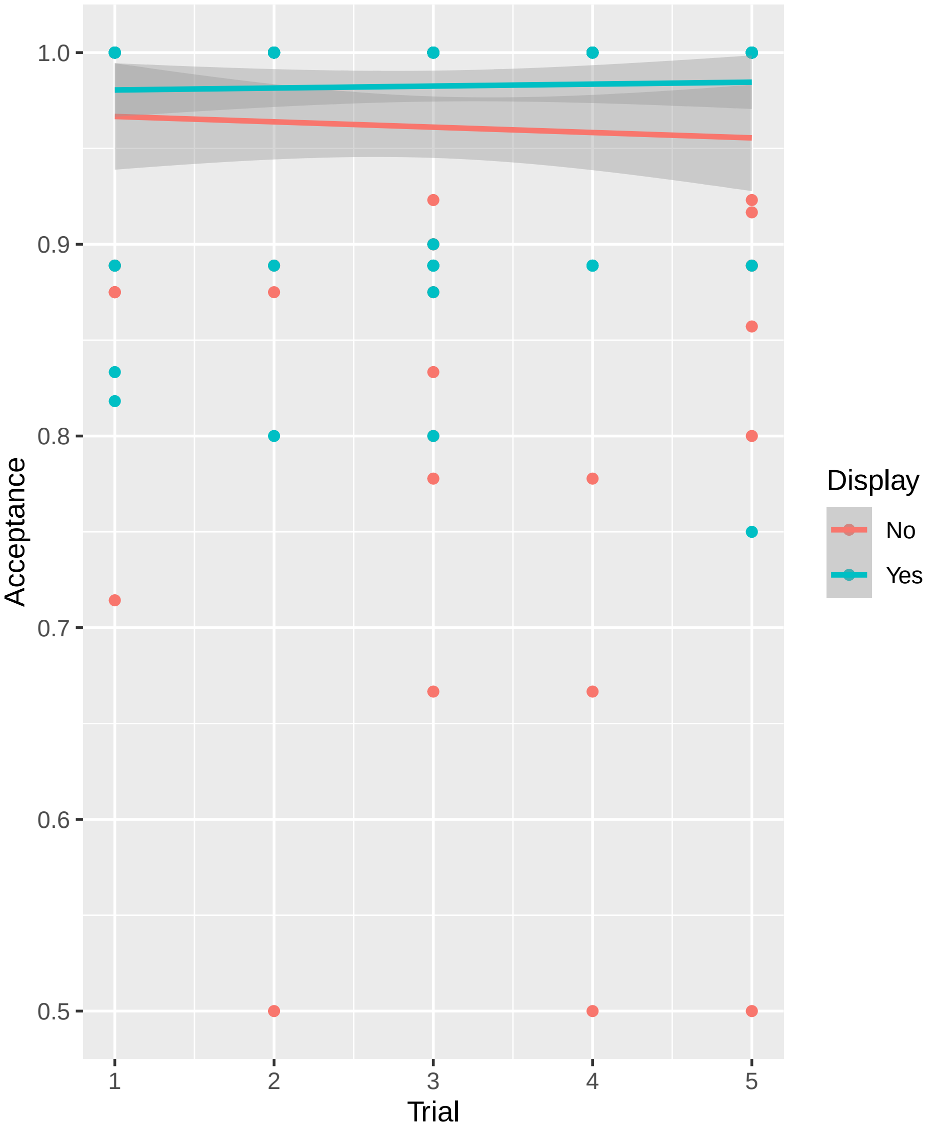


Figure 6. Acceptance of recommendations over time with and without the option-centric rationale display.

4.1 Performance over time with and without option-centric rationale display

Neither the display ($t(1, 250) = 2.59, p = .01$) or time ($t(1, 250) = 2.59, p = .01$) had a significant effect on the performance score.

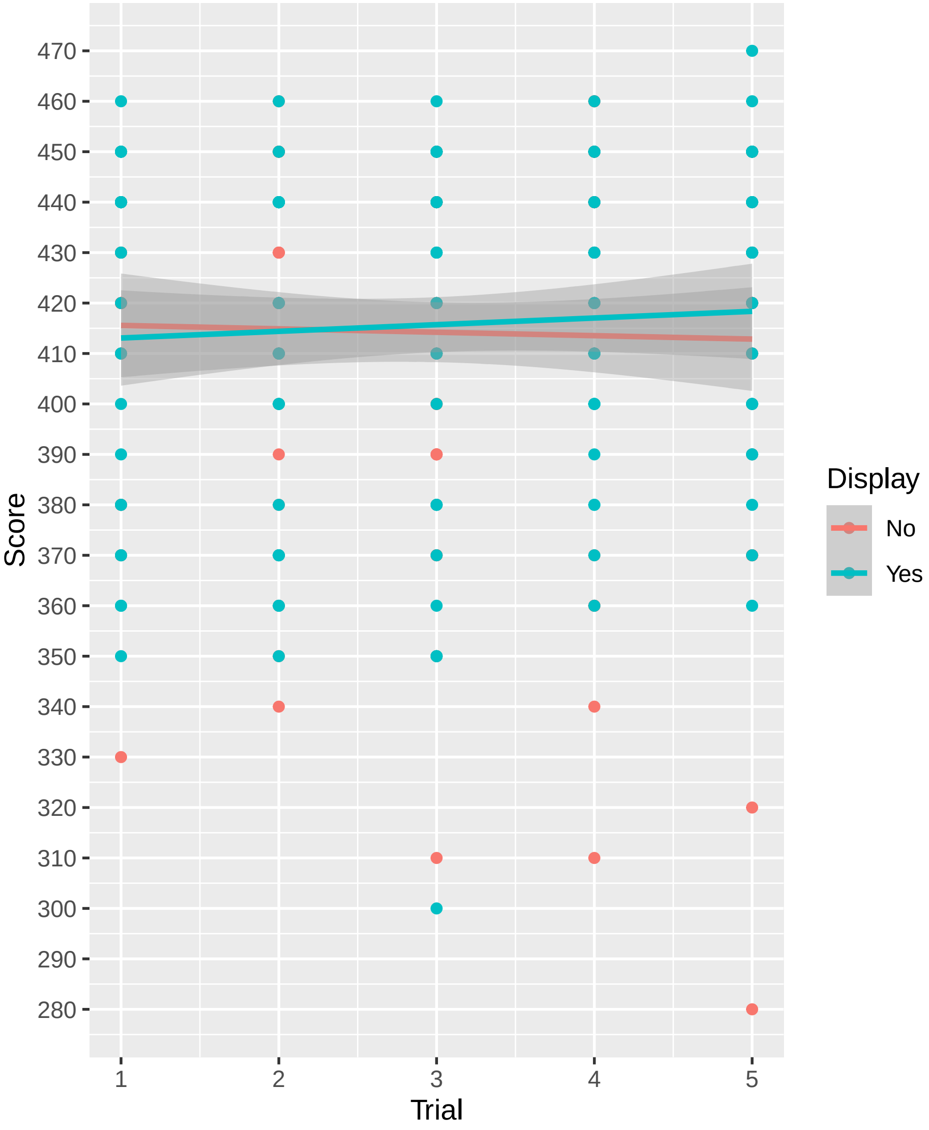


Figure 7. Human-autonomy team performance over time with and without the option-centric rationale display.

5. DISCUSSION

In the present study, we examine how autonomy transparency affects people's trust, dependence behavior, and human-autonomy team performance *over time*. Autonomy transparency is achieved through the option-centric rationale display that explicitly explores the option space (i.e., all the possible options/actions that an autonomy could take on) and presents the rationale why a particular option is the most appropriate by detailing all the available next locations and the criteria for recommending a particular location.

Concerning people's trust in the intelligent agent, we find that people have higher trust in the agent with enhanced transparency provided by the option-centric rationale display, in line with prior literature (Chen & Barnes, 2014; Lyons & Havig, 2014). More importantly, our results reveal that with the display, people's trust in the intelligent agent increases gradually; Without the display, trust nearly stays unchanged. The result seems to contradict prior research showing that as people gained more experience interacting with an automated aid, their trust in the aid increased regardless of the absence/presence of more transparency information (Yang et al., 2017). In the study of Yang et al. (2017), half of the participants were presented with the binary alerts from an automated decision aid (i.e., aid recommends whether or not there is a threat) and the other half the likelihood alerts (i.e., aid recommends whether or not there is a threat and shows the confidence level of the recommendation). Trust increments were observed in both the binary and likelihood conditions. This inconsistent result could have been due to the different tasks employed in the studies. In the study of Yang et al. (2017), the task was threat detection with the help of an imperfect automated threat detector, which was pretty simple compared to the Treasure Hunter task where the participant and the intelligent agent needed to perform cognitively demanding inferences. In such a demanding task, without the enhanced transparency though the display, participants may have increasing difficulty figuring out the strategy of the intelligent agent, leading to stagnant trust over time.

It is worth emphasizing the difference between trust and trust calibration, which are related but different constructs. Trust in autonomy is a person's "attitude that an (autonomous) agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability", whereas trust-calibration refers to the correspondence between a person's trust and the autonomy's actual capability (Lee & See, 2004). High trust in an incapable autonomy is unjustified and will harm instead of

TABLE 3: *Mean and Standard Error (SE) values of the Intelligent Assistant's Score and the Optimal Score for Each Test Map*

Test Map ID	1	2	3	4	5	6	7	8	9	10
Intelligent Assistant's Score	450 ± 0	381 ± 2.4	430.5 ± 0.5	440 ± 0	415 ± 1.1	421 ± 0.7	363.5 ± 3.5	386.5 ± 3.4	428.5 ± 1.7	447.5 ± 1.9
Optimal Score	470	460	460	460	450	450	450	440	460	470
Ratio (%)	95.7	82.8	93.6	95.7	92.2	93.6	80.8	87.8	93.2	95.2

benefit performance. In our present study, the intelligent agent was near-optimal. Table 3 details the optimal score that an omniscient agent could obtain and the score that the knowledge based intelligent assistant used in the present study obtained. The optimal score was calculated assuming that the intelligent assistant was omniscient (i.e., the map was known to the intelligent assistant). The intelligent assistant's score was calculated by having the autonomous agent play the treasure hunter game by itself for 20 times. As shown in Table 3, the intelligent assistant's performance was close to the optimal score. The ratio between the intelligent assistant's score and the optimal score was on average 91.1%. As shown in Figure 5, the final trust score in the intelligent agent was on average around 8 with the option-centric rationale display and around 6.8 without the display on a 9-point Likert scale. In this case, greater trust (but not yet over-trusting) largely indicates better trust-calibration.

With respect to dependence behaviors, consistent with findings from previous studies (Beller, Heesen, & Vollrath, 2013; Forster, Naujoks, Neukum, & Huestegge, 2017; Koo et al., 2014, 2016), we find that enhanced transparency led to higher acceptance. However, there was a non-significant effect of time. This lack of significance could have been due to a ceiling effect. As Figure 6 shows, the acceptance of recommendations made by the intelligent agent was near 100% throughout the five trials.

To our surprise, neither the option-centric rationale display nor time had an effect on the performance score. The lack of significance could have been due to two reasons. First, similar to the discussion on participants' dependence behaviors, there could be a ceiling effect. Second, as described in the Method section, the testbed is inherently highly stochastic, meaning that keeping following the recommendations from the "on average" competent agent could lead to unfavorable outcomes.

Although we only tested the option-centric rationale display on a simulated game with a small action space, the display can be applied to other decision-making agents with a larger action space, for instance, an epsilon-greedy agent with finite (i.e., countable) action space. The epsilon-greedy agent balances exploration and exploitation by choosing the optimal action some times and the exploratory action other times. The exploratory action is not the optimal action at a particular step. However, by further exploring the environment, the agent can obtain higher rewards in the subsequent steps and higher accumulative rewards. The option-centric rationale display can list all possible actions with the expected reward and the number of times the optimal/exploratory action has been taken to indicate the necessity of exploring the environment. For a large action space, the display can present a subspace of the action space that contains the optimal and near-optimal actions by listing the actions with top expected scores. The other (far from optimal) actions can be displayed if requested. Further research is needed to determine the size of subspace to be displayed.

6. CONCLUSION

The advance in artificial intelligence and machine learning empowers a new generation of autonomous systems. However, human agents increasingly have difficulty deciphering autonomy-generated solutions. The lack of transparency contributes to the lack of trust in autonomy and suboptimal human-autonomy team performance (Chen & Barnes, 2014; de Visser et al., 2018; Endsley, 2017; Lyons & Havig, 2014; Lyons et al., 2016; Yang et al., 2017). In this study, we proposed an option-centric rationale display for enhancing autonomy transparency. The display details all the potential actions and the criteria for choosing a particular action, and highlights the final recommendation. The results indicate that the presence/absence of the display significantly affected people's trust evolution over time. With the display, their trust increased significantly and became more calibrated over time.

The results should be reviewed in light of several limitations. First, the intelligent assistant used in the present study was highly capable. However, in the real world, an intelligent assistant could be less capable in situations of high uncertainty and ambiguity. Further research with less capable autonomous agents is needed to validate the generalization of the display. Second, the action space in the simulated game was limited. We discussed the application of the option-centric rationale display on domains with larger action spac in the results section. Further research is needed to examine the proposed solutions. Third, similar to a few previous studies (Manzey, Reichenbach, & Onnasch, 2012; Yang et al., 2017), we used a one-item scale to measure trust. The one-item scale could fail to capture all of the sub-dimensions of trust compared to multi-dimension scales such as the 12-item trust scale in Jian, Bisantz, and Drury (2000). Further research should investigate the possibility of developing a succinct multi-item trust scale that can be used in querying trust repeatedly.

References

Bagheri, N., & Jamieson, G. A. (2004). The impact of context-related reliability on automation failure detection and scanning behaviour. In *2004 IEEE International Conference on Systems, Man and Cybernetics* (pp. 212–217). IEEE.

Beller, J., Heesen, M., & Vollrath, M. (2013). Improving the Driver–Automation Interaction: An Approach Using Automation Uncertainty. *Human Factors*, *55*(6), 1130–1141.

Chen, J. Y. C., & Barnes, M. J. (2014). Human-Agent teaming for multirobot control: A review of human factors issues. *IEEE Transactions on Human-Machine Systems*, *44*(1), 13–29.

Chen, J. Y. C., Lakhmani, S. G., Stowers, K., Selkowitz, A. R., Wright, J. L., & Barnes, M. (2018). Situation awareness-based agent transparency and human-autonomy teaming effectiveness. *Theoretical Issues in Ergonomics Science*, *19*(3), 259–282.

Chen, J. Y. C., Procci, K., Boyce, M., Wright, J., Garcia, A., & Barnes, M. J. (2014). *Situation awareness-based agent transparency (No. ARL-TR-6905)* (Tech. Rep.). Aberdeen Proving Ground, MD.

de Visser, E. J., Pak, R., & Shaw, T. H. (2018). From 'automation' to 'autonomy': the importance of trust repair in human–machine interaction. *Ergonomics*, *133*(1), 1–19.

de Visser, E. J., Peeters, M. M., Jung, M. F., Kohn, S., Shaw, T. H., Pak, R., & Neerincx, M. A. (2020). Towards a theory of longitudinal trust calibration in human–robot teams. *International Journal of Social Robotics*, *12*(2), 459–478.

Dixon, W. (1953). Processing data for outliers. *Biometrics*, *9*(1), 74–89.

Du, N., Haspiel, J., Zhang, Q., Tilbury, D., Pradhan, A. K., Yang, X. J., & Robert, L. P. (2019). Look who's talking now: Implications of AV's explanations on driver's trust, AV preference, anxiety and mental workload. *Transportation Research Part C: Emerging Technologies*, *104*, 428–442.

Du, N., Huang, K. Y., & Yang, X. J. (2020). Not All Information Is Equal: Effects of Disclosing Different Types of Likelihood Information on Trust, Compliance and

Reliance, and Task Performance in Human-Automation Teaming. *Human Factors*, *62*(6), 987–1001.

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies*, *58*(6), 697–718.

Endsley, M. R. (2017). From Here to Autonomy: Lessons Learned From Human–Automation Research. *Human Factors*, *59*(1), 5–27.

Fletcher, K. I., Bartlett, M. L., Cockshell, S. J., & McCarley, J. S. (2017). Visualizing probability of detection to aid sonar operator performance. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 61, pp. 302–306).

Forster, Y., Naujoks, F., Neukum, A., & Huestegge, L. (2017, December). Driver compliance to take-over requests with different auditory outputs in conditional automation. *Accident Analysis and Prevention*, *109*, 18–28.

Guo, Y., & Yang, X. J. (2021). Modeling and predicting trust dynamics in human–robot teaming: A bayesian inference approach. *International Journal of Social Robotics*, *13*, 1899–1909.

Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, *4*(1), 53–71.

Koo, J., Kwac, J., Ju, W., Steinert, M., Leifer, L., & Nass, C. (2014). Why did my car just do that? Explaining semi-autonomous driving actions to improve driver understanding, trust, and performance. *International Journal on Interactive Design and Manufacturing (IJIDeM)*, *9*(4), 269–275.

Koo, J., Shin, D., Steinert, M., & Leifer, L. (2016). Understanding driver responses to voice alerts of autonomous car operations. *International Journal of Vehicle Design*, *70*(4), 377–17.

Lee, J. D., & See, K. A. (2004). Trust in automation: designing for appropriate reliance. *Human Factors*, *46*(1), 50–80.

Lyons, J. B., & Havig, P. R. (2014). Transparency in a Human-Machine Context:

Approaches for Fostering Shared Awareness/Intent. In R. Shumaker & S. Lackey (Eds.), *Virtual, augmented and mixed reality. designing and developing virtual and augmented environments* (pp. 181–190). Springer International Publishing.

Lyons, J. B., Ho, N. T., Koltai, K. S., Masequesmay, G., Skoog, M., Cacanindin, A., & Johnson, W. W. (2016). Trust-Based Analysis of an Air Force Collision Avoidance System. *Ergonomics in Design*, *24*(1), 9–12.

MacLean, A., Young, R. M., Bellotti, V. M. E., & Moran, T. P. (1991). Questions, options, and criteria: Elements of design space analysis. *Humman-Computer Interaction*, *6*(3), 201–250.

Manzey, D., Reichenbach, J., & Onnasch, L. (2012, February). Human Performance Consequences of Automated Decision Aids: The Impact of Degree of Automation and System Experience. *Journal of Cognitive Engineering and Decision Making*, *6*(1), 57–87.

McGuirl, J. M., & Sarter, N. B. (2006). Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. *Human Factors*, *48*(4), 656–665.

Mercado, J. E., Rupp, M. A., Chen, J. Y. C., Barnes, M. J., Barber, D., & Procci, K. (2016). Intelligent agent transparency in human–agent teaming for multi-uxv management. *Human Factors*, *58*(3), 401-415.

Miller, C. A. (2018). Displaced interactions in human-automation relationships: Transparency over time. In D. Harris (Ed.), *Engineering Psychology and Cognitive Ergonomics* (p. 191-203). Cham: Springer International Publishing.

Russell, S., & Norvig, P. (2010). *Artificial Intelligence: A Modern Approach* (3rd ed.). Pearson.

Seong, Y., & Bisantz, A. M. (2008, July). The impact of cognitive feedback on judgment performance and trust with decision aids. *International Journal of Industrial Ergonomics*, *38*(7-8), 608–625.

Sheridan, T. B., & Verplank, W. L. (1978). *Human and computer control of undersea teleoperators* (Tech. Rep.). Cambridge, MA: Man Machine Systems Laboratory, MIT.

Theodorou, A., Wortham, R. H., & Bryson, J. J. (2017). Designing and implementing transparency for real time inspection of autonomous robots. *Connection Science*, *29*(3), 230–241.

Walliser, J. C., de Visser, E. J., & Shaw, T. H. (2016). Application of a system-wide trust strategy when supervising multiple autonomous agents. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 60, pp. 133–137).

Wang, L., Jamieson, G. A., & Hollands, J. G. (2009). Trust and reliance on an automated combat identification system. *Human Factors*, *51*(3), 281-291.

Yang, X. J., Schemanske, C., & Searle, C. (2021). Toward Quantifying Trust Dynamics: How People Adjust Their Trust After Moment-to-Moment Interaction With Automation. *Human Factors*, 00187208211034716.

Yang, X. J., Unhelkar, V. V., Li, K., & Shah, J. A. (2017). Evaluating effects of user experience and system transparency on trust in automation. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction (HRI '17)* (pp. 408–416). New York, NY, USA: ACM.

Biographies

Ruikun Luo is a Ph.D. candidate at the Robotics Institute, University of Michigan, Ann Arbor. Prior to joining the University of Michigan, he obtained a M.S. in Mechanical Engineering from Carnegie Mellon University in 2014 and a B.S. in Mechanical Engineering and Automation from Tsinghua University, China in 2012.

Na Du is an Assistant Professor in the Department of Informatics and Networked Systems at the University of Pittsburgh. She received her Ph.D. degree in Industrial & Operations Engineering from the University of Michigan and Bachelor's degree in Psychology from Zhejiang University. Her research interests include human-computer interaction, transportation human factors, computational modeling of human behaviors, and human-centered design.

X. Jessie Yang is an Assistant Professor in the Department of Industrial and Operations Engineering and an affiliated faculty at the Robotics Institute, University of Michigan, Ann Arbor. She obtained a PhD in Mechanical and Aerospace Engineering (Human Factors) from Nanyang Technological University, Singapore in 2014.