

# Enhancing autonomy transparency: an option-centric rationale approach

Ruikun Luo

Robotics Institute, University of Michigan, Ann Arbor

Na Du

Industrial and Operations Engineering, University of Michigan, Ann Arbor

X. Jessie Yang

Industrial and Operations Engineering, University of Michigan, Ann Arbor

**Manuscript type:** *Original Research*

**Running head:** Enhancing autonomy transparency

**Word count:** 4200

**Corresponding author:** X. Jessie Yang, 1205 Beal Avenue, Ann Arbor, MI48109,

Email: xijyang@umich.edu

**Acknowledgement:** We would like to thank Kevin Y. Huang for his assistance in data collection.

## Abstract

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: Humans 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 sub-optimal team performance. To enhance autonomy transparency, this study proposed an *option-centric rationale* display and evaluated its effectiveness. We developed a game *Treasure Hunter* wherein a human uncovers a map for treasures with the help from an intelligent assistant, and conducted a human-in-the-loop experiment with 34 participants. Results indicated that by conveying the intelligent assistant's decision-making rationale via the *option-centric rationale* display, participants had higher trust in the system and calibrated their trust faster. Additionally, higher trust led to higher acceptance of recommendations from the intelligent assistant, and in turn higher task performance.

**Keyword:** Transparent autonomy, Transparent automation, Design rationale, Trust in automation, Trust calibration, Propositional logic.

## 1. INTRODUCTION

While the advances in artificial intelligence and machine learning empowers 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 perceived autonomy as a mysterious black box. The lack of transparency contributes to the lack of trust in autonomy and sub-optimal 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).

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 fairly 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 human operators calibrate their trust and enhance human-automation team performance (McGuirl & Sarter, 2006; Walliser, de Visser, & Shaw, 2016; Wang, Jamieson, & Hollands, 2009). Other studies reported that human operators did not trust or depend on automated decision aids appropriately even when the likelihood information was disclosed (Bagheri & Jamieson, 2004; Fletcher,

Bartlett, Cockshell, & McCarley, 2017). Recently, Du, Huang, and Yang (in press) proposed a framework for reconcile the mixed results and showed that not all likelihood information is equal in aiding human-autonomy team performance. Presenting the predictive values and the overall success likelihood is more beneficial than presenting the hit and correct rejection rates.

Another body of research has investigated the impact of providing hand-crafted explanations of autonomy's behaviors. For example, Dzindolet, Peterson, Pomranky, Pierce, and Beck (2003) showed that providing hand-drafted explanations of automation failures can lead to increased trust. The studies of Koo et al. (2014) and Koo, Shin, Steinert, and Leifer (2016) showed that by informing the drivers of the reasons for automated braking (e.g. road hazard ahead) decreased drivers' anxiety, 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 gone beyond either conveying likelihood information or relying on hand-crafted explanation, and has formally defined and examined autonomy transparency. Notably, Mercado et al. (2016) proposed the situation awareness (SA)-based agent transparency model to convey information supporting the human agent's perception, comprehension, and projection of an intelligent assistant's recommendations.

In this study, we wished to propose the *option-centric rationale* approach for enhancing autonomy transparency. Inspired by the research on design rationale, the option-centric rationale approach 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 stake holders 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.

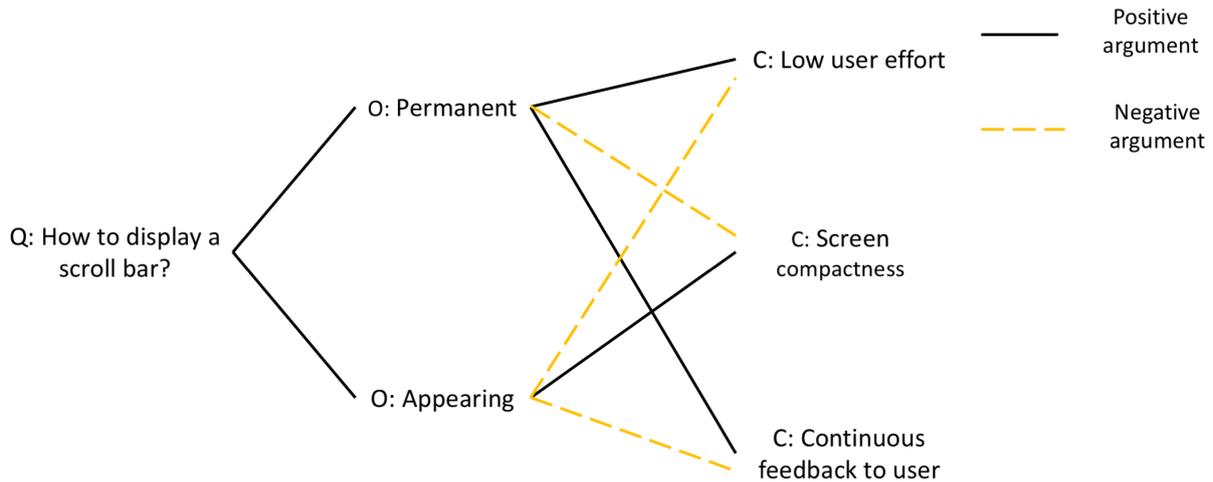


Figure 1. An illustration of the QOC notation. Adapted from MacLean et al. (1991).

One major approach to representing design rationale is design space analysis. It uses a semi-formal notation called QOC (Questions, Options, and Criteria): *Questions* identify key design issues, *Options* provide possible answers to the Questions, and *Criteria* illustrate the criteria used for assessing and comparing the Options. The QOC notation creates an explicit representation of a structured space of design alternatives and the consideration for choosing among the different choices. It has been widely applied in the design of human-machine interface (MacLean et al., 1991; Oulasvirta, Feit, Lähteenlahti, & Karrenbauer, 2017; Sutcliffe et al., 2018). Figure 1 illustrates the QOC representation of the design space of the scroll bar mechanism in the Xerox Common Lisp (XCL) environment. The design question of interest is how to display the scroll bar. There are two options: the scroll bar appears permanently or only appears when the cursor is moved over the edge of a window. Choosing among various options requires a range of considerations and reasoning over those considerations. For the scroll bar example, the reasoning criteria includes low user effort, screen compactness and continuous feedback to the user. The QOC notation also provides a way to represent an *Assessment* of whether an option does or does not satisfy a criterion. The “permanent”

option ensures continuous feedback and low user effort at the cost of screen space.

## 2. THE PRESENT STUDY

In the present study, we proposed the *option-centric rationale* display for enhancing autonomy transparency, and evaluated its effectiveness via a human-in-the-loop experiment. In the experiment, a human operator uncovered a map for treasures with the help from an intelligent assistant. The intelligent assistant's decision-making rationale are conveyed in the *option-centric rationale* display. We tested the following hypotheses:

First, the *option-centric rationale* display explicitly explores the option space (i.e. all the possible options/actions that an autonomy could take on) and present the rationale why a particular option is the most appropriate at a given context. We expected that the enhanced transparency will lead to higher trust:

**H1:** When the *option-centric rationale* display is present, human agents will have higher trust in the autonomous agent.

Second, trust has been recognized as one, if not the most, crucial factor determining the use of automated or autonomous technologies (Hoff & Bashir, 2015). Trust in automation is defined as the "attitude that an agent will help achieve an individual's goals in situations characterized by uncertainty and vulnerability (Lee & See, 2004)". The underlying idea is that a human agent's decision whether or not to use an automated or autonomous technology depends on the extent to which he or she trusts it. When trust is high, the human agent is more likely to use the technology. Besides the research focus on trust, several researchers have proposed that the human agent's dependence behavior is inversely related to his or her self-confidence in performing the task by themselves manually without the assistance of the intelligent agent (de Vries, Midden, & Bouwhuis, 2003; Kantowitz, Hanowski, & Kantowitz, 1997; Moray, Inagaki, & Itoh, 2000). Therefore, we hypothesized that:

**H2:** Trust and self-confidence affect people's acceptance behaviors. Higher trust will lead to higher acceptance and higher self-confidence will reduce acceptance.

Third, Parasuraman and Riley (1997) categorized different types of automation use and showed disusing a reliable automation harmed task performance. In the present study, the intelligent assistant is a knowledge-based agent and reasons using propositional logic. Therefore, we hypothesized that:

**H3:** Given a highly capable autonomous agent, higher acceptance leads to higher performance.

Last, most existing studies on trust in automation, or more recently, trust in autonomy, measured trust once at the end of an experiment via a questionnaire (a "snapshot" of trust). Only a limited number of studies have viewed trust as a dynamic variable that can strength or decay over time (Lee & Moray, 1992; Manzey, Reichenbach, & Onnasch, 2012; Yang et al., 2017; Yang, Wickens, & Hölttä-Otto, 2016). Prior studies showed that human agents calibrate their trust based on their experience interacting with the automated or autonomous technology (Yang et al., 2017). With enhanced autonomy transparency, we expected to see a faster trust calibration process:

**H4:** People adjust their trust in the autonomous agent as they gain more experience working with the autonomous agent. In particular, with the *option-centric rationale* display, the rate of adjustment will be faster.

### 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 2a & 2b). Each step, the hunter can move to an unvisited location which is connected to the visited locations. Figure 2b 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 2a) or encounter a wumpus (shown in B3 in Figure 2a). 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 2a) and a breeze when a pit is in an adjacent location (shown as B1, C2, D1 in Figure 2a). There is one and only one gold bar/wumpus in 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.

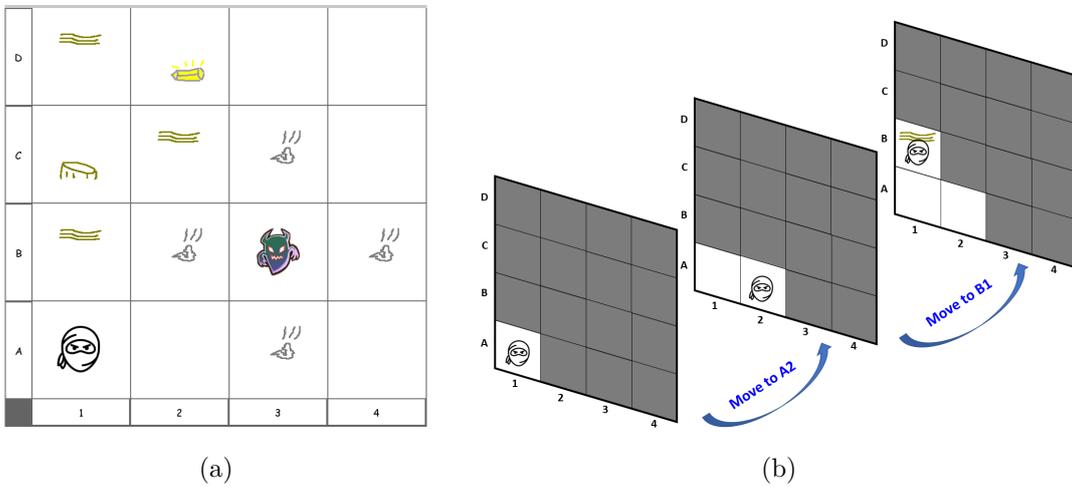


Figure 2. (a) An example map in *Treasure Hunter*. Each square is denoted by the row number (from 1 to 4) and the column number (from A to D). (b) First two steps of a hunter moving in the map.

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 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 wumups	-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  $D_2$ ): (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 3: 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 3 are the impossible cases because the pit and wumpus cannot co-exist in one location. For each case in Figure 3, 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.

		<b>Pit</b>		
		<b>Y</b>	<b>N</b>	<b>NA</b>
<b>Wumpus</b>	<b>Y</b>		<b>1</b>	
	<b>N</b>	<b>2</b>	<b>3</b>	<b>4</b>
	<b>NA</b>		<b>5</b>	<b>6</b>

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 3 & 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.

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 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) wherein the autonomy narrows down the action selections and suggests one alternative to the human operator. Similar experimental paradigms have been used to examine trust in and dependence on automation (Manzey et al., 2012; McBride, Rogers, & Fisk, 2011; Neyedli, Hollands, & Jamieson, 2011; Wang et al., 2009).

Figure 4 shows the *option-centric rationale* display proposed in this study. The *option-centric rationale* 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

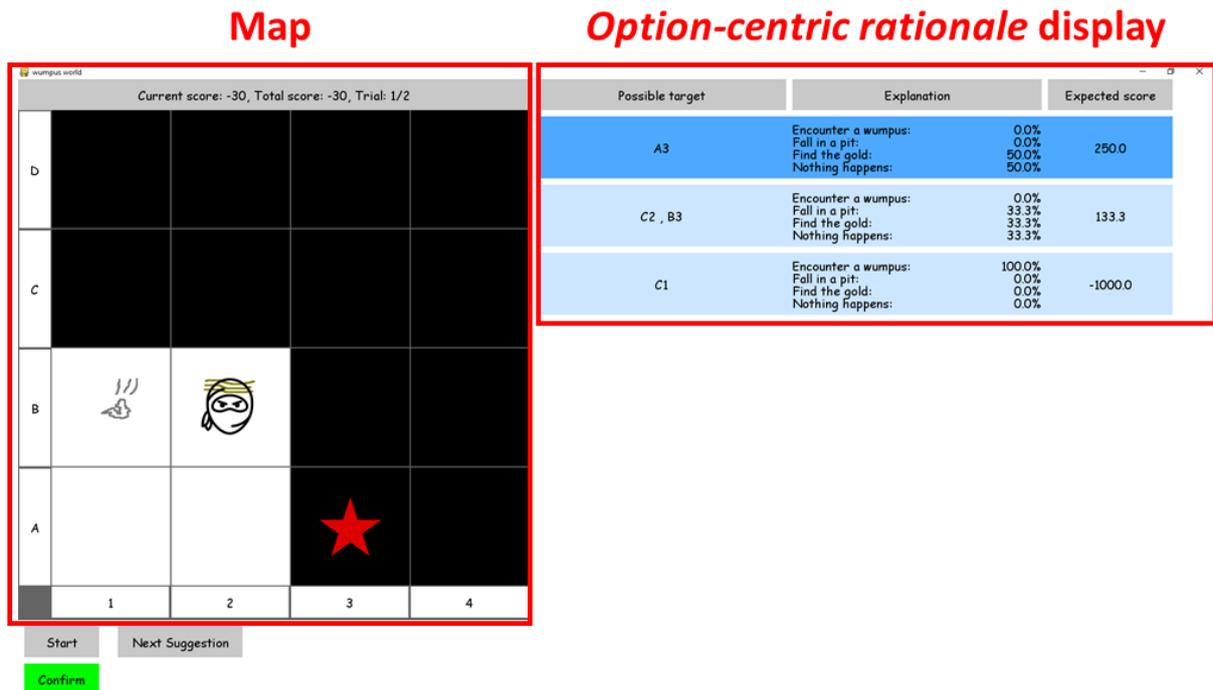


Figure 4. Testbed with *option-centric rationale* display.

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 variable in the experiment was the presence/absence of the *option-centric rationale* display. The order of the two conditions was counterbalanced to eliminate potential order effects. In each condition, participants played the game on 5 different maps. In the absence of the *option-centric rationale* display condition, the participant only saw a red star that indicated the recommendation by the intelligent assistant.

### 3.4 Measures

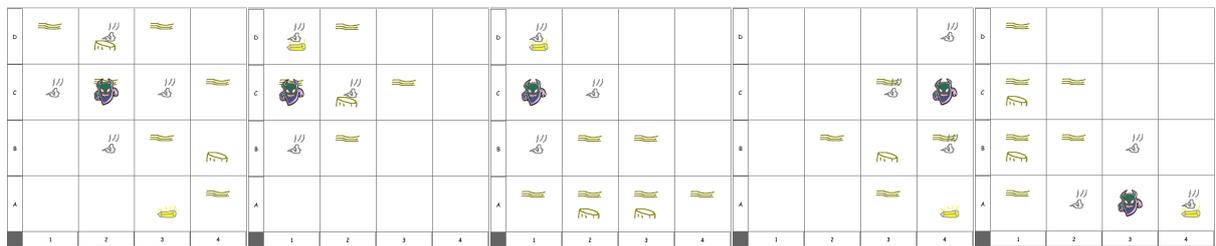
We measured three groups of dependent variables: subjective responses, behavioral responses and performance. After completing each map, participants were asked to report their trust in the intelligent assistant and their self-confidence to accomplish the task without the intelligent assistant using two 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

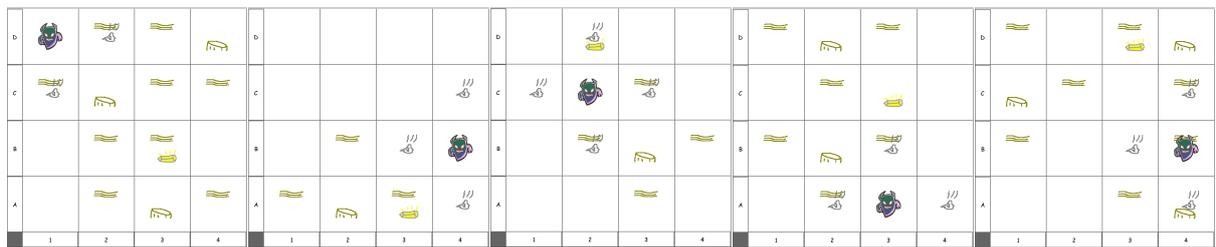
In order to eliminate the inherent randomness of the task, we carefully selected the maps used in the experiment (Figure 5). First, we randomly generated 100 maps and ran the game only with the intelligent assistant 20 times for each map (i.e, always accepted the recommendations from the intelligent 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 which fulfilled 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

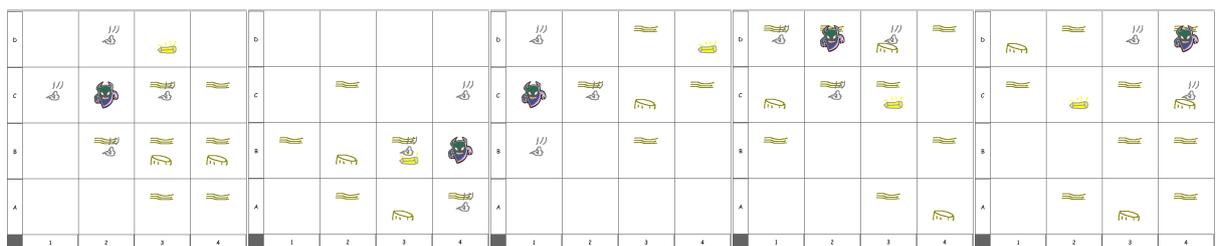
onward, participants played the game with the help of the intelligent assistant. The second and fourth practice maps were similar to the maps participants experienced in the real test. The third and the fifth maps contained a pit next to the start location. The reason for selecting the two maps (i.e. the third and the fifth map) was to help participants fully understand the stochastic nature of the game. For example, in the fifth training map (Figure 5), 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.



(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 5. 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.

### 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. 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.

## 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 hypotheses were tested using data from the remaining 28 participants (Mean age = 21.25 years, SD = 1.72 years). To analyze the data, we performed structural equation modeling (SEM) using SPSS AMOS. SEM is a multivariate statistical method widely used in behavioral science (Bollen & Noble, 2011; Goldberger, 1972). It uses a path diagram to specify the model, which indicates the relationships between the variables. An arrow represents an effect of a variable on another. In SEM, multiple regressions are simultaneously estimated to indicate the strength of each relationship (arrow). Measurement of fit in SEM provides information about how well the model fits the data. A good-fit model is required before interpreting the causal paths of the structural model.

### 4.1 Relationships between the *option-centric rationale* display, trust, self-confidence, recommendation acceptance and performance

To test hypotheses 1, 2 and 3, we constructed and tested a model as specified in Figure 6. Based on prior research (Hooper, Coughlan, & Mullen, 2008; Hu & Bentler,

1999; Steiger, 2007), multiple model fit indices were used to check the model fit. Our model demonstrated a good fit, with  $\chi^2(6) = 3.776, p = .71$ , RMSEA = 0.00, CFI = 1.00, SRMR = .058 .

The model is shown in Figure 6, with annotated standardized weights and significance (noted with asterisks). Standardized weights can be directly compared to assess the strength of each factor’s impact on the subsequent factor. In our study, the chains connecting display type, trust, recommendation acceptance, and task performance were significant: the *option-centric rationale* display increased participants’ trust, which significantly affected their recommendation acceptance, and in turn impacted the task performance. Yet, participants’ self-confidence did not affect their recommendation acceptance.

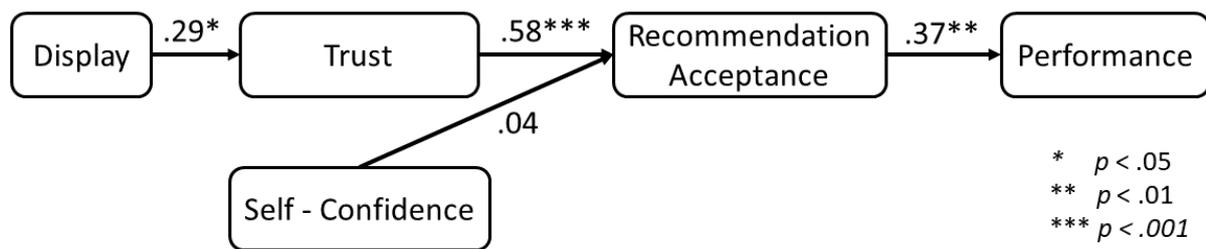


Figure 6. The structural equation model indicates the relationship between the display, trust, self-confidence, acceptance and task performance.

## 4.2 Trust adjustment overtime

To examine how trust changes overtime, we conducted a linear mixed effects analysis, with the presence/absence of the *option-centric rationale* display and trial number as fixed effects and intercepts for subjects as random effects. Results are reported as significant for  $\alpha < .05$ .

There was a significant interaction effect between display and time,  $F(1, 249) = 5.675, p = .018$  (See Figure 7). Specifically, when participants were provided with the *option-centric rationale* display, their trust in the autonomous agent increased as they gained more experiences,  $F(1, 111) = 43.714, p < .001$ . When the display was not present, however, their trust did not change significantly,  $F(1, 111) = 0.549, p = 0.460$ .

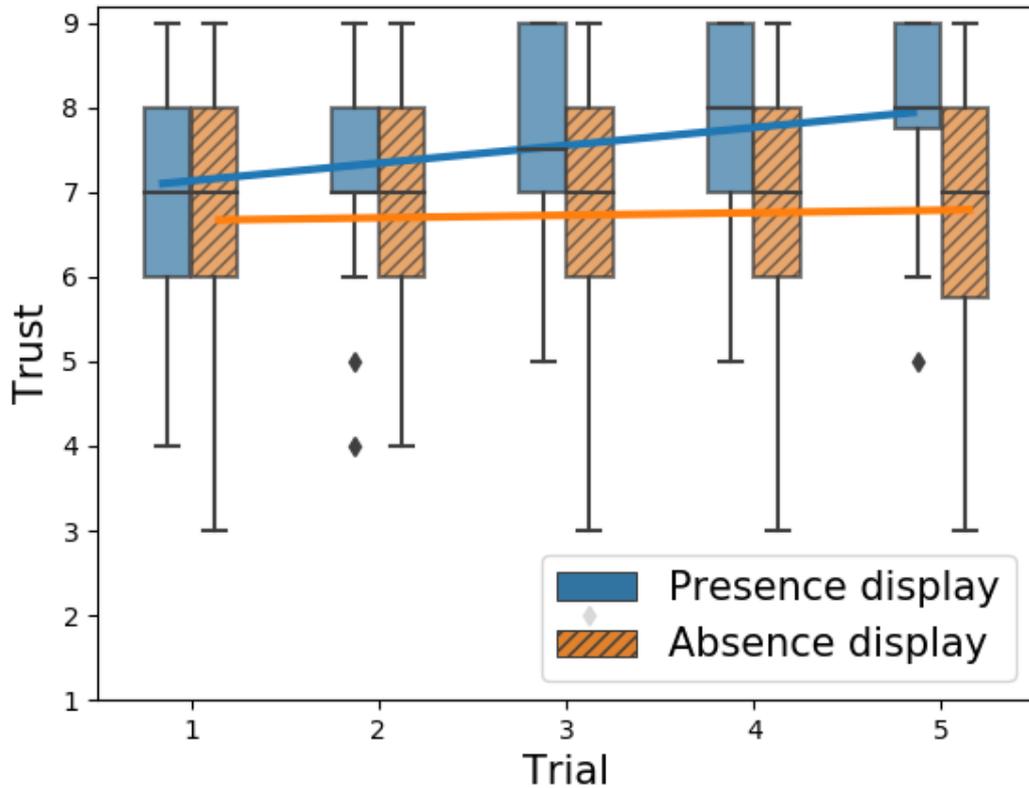


Figure 7. The change of trust over time.

## 5. DISCUSSION

In this study, we proposed an *option-centric rationale* display. The display explicitly conveys the intelligent assistant’s decision-making rationale by detailing all the available next locations and the criteria for recommending a particular location.

Our results indicate that with the *option-centric rationale* display, human operators trusted the intelligent assistant more. In addition, consistent with findings from previous studies (Beller, Heesen, & Vollrath, 2013; Forster, Naujoks, Neukum, & Huestegge, 2017; Koo et al., 2014, 2016), we also found that trust significantly affected people’s acceptance behaviors. As human operators’ trust in the autonomous agent increased, they accepted more recommendations provided by the intelligent assistant. Meanwhile, human operators’ self-confidence did not influence their recommendation acceptance, which was in line with previous research (Lee & Moray, 1994; Moray et al., 2000).

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

We found that higher recommendation acceptance led to higher task performance. We argued that this positive relationship hinged on the capability of the intelligent assistant, which was near-optimal in our study. 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 could obtain. 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. 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%.

Prior research showed that as people gained more experience interacting with the autonomous agent, their trust in the agent will evolve and eventually stabilize if enough interaction is warranted (Yang et al., 2017). Results from this study supported prior results. More importantly, human operators adjusted their trust faster when the *option-centric rationale* display was presented. This implies that human operators would require less amount of interaction with the autonomous agent to calibrate their trust before trust reaching a steady state. This benefit is also attributable to the enhanced transparency provided by the *option-centric rationale* display. The list of all available options prevents severe drop of trust when optimal performance was not achieved simply because of the inherent randomness in the game.

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 *option-centric rationale* 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 and therefore perceive autonomy as a mysterious black box. The lack of transparency contributes to the lack of trust in autonomy and sub-optimal 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 *option-centric rationale* 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 in the autonomous agent and human operators' trust increased faster when the display was provided. 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 space. Further research is needed to examine the proposed solutions.

## 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, March). Improving the Driver–Automation Interaction: An Approach Using Automation Uncertainty. *Human Factors*, *55*(6), 1130–1141.
- Bollen, K. A., & Noble, M. D. (2011, oct). Structural equation models and the quantification of behavior. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(SUPPL. 3), 15639–15646.
- Chen, J. Y. C., & Barnes, M. J. (2014, February). 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., 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, April). From ‘automation’ to ‘autonomy’: the importance of trust repair in human–machine interaction. *Ergonomics*, *133*(1), 1–19.
- de Vries, P., Midden, C., & Bouwhuis, D. (2003, jun). The effects of errors on system trust, self-confidence, and the allocation of control in route planning. *International Journal of Human-Computer Studies*, *58*(6), 719–735.
- 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. (in press). 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*, 0018720819862916.
- Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003, June). 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.
- Goldberger, A. S. (1972). Structural Equation Methods in the Social Sciences. *Econometrica*, 40(6), 979–1001.
- Hoff, K. A., & Bashir, M. (2015, April). Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors*, 57(3), 407–434.
- Hooper, D., Coughlan, J., & Mullen, M. (2008). Structural equation modelling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods*, 6, 53-60.
- Hu, L.-t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal*, 6(1), 1–55.
- Kantowitz, B. H., Hanowski, R. J., & Kantowitz, S. C. (1997). Driver reliability requirements for traffic advisory information. In *Ergonomics and safety of intelligent driver interfaces* (p. 1-22). Lawrence Erlbaum Associates.
- Koo, J., Kwac, J., Ju, W., Steinert, M., Leifer, L., & Nass, C. (2014, April). 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., & Moray, N. (1992). Trust control strategies and allocation of function in human-machine systems. *Ergonomics*, 35(10), 1243–1270.
- Lee, J., & Moray, N. (1994). Trust, self-confidence and operators' adaptation to automation. *International Journal of Human-Computer Studies*, 40(1), 153–184.
- Lee, J., & See, K. (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. doi: 10.1177/1064804615611274
- MacLean, A., Young, R. M., Bellotti, V. M. E., & Moran, T. P. (1991, September). Questions, options, and criteria: Elements of design space analysis. *Human-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.
- McBride, S. E., Rogers, W. A., & Fisk, A. D. (2011, nov). Understanding the effect of workload on automation use for younger and older adults. *Human Factors*, 53(6), 672–686.
- 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.
- Moray, N., Inagaki, T., & Itoh, M. (2000). Adaptive automation, trust, and self-confidence in fault management of time-critical tasks. *Journal of Experimental Psychology: Applied*, 6(1), 44.
- Neyedli, H. F., Hollands, J. G., & Jamieson, G. A. (2011). Beyond identity: incorporating system reliability information into an automated combat identification system. *Human Factors*, 53(4), 338–355.
- Oulasvirta, A., Feit, A., Lähteenlahti, P., & Karrenbauer, A. (2017). Computational support for functionality selection in interaction design. *ACM Transactions on Computer-Human Interaction*, 24(5), 34:1–34:30.
- Parasuraman, R., & Riley, V. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors*, 39(2), 230–253.
- Russell, S., & Norvig, P. (2010). *Artificial Intelligence: A Modern Approach* (3rd ed.). Pearson.
- Sheridan, T. B., & Verplank, W. L. (1978). *Human and computer control of undersea teleoperators* (Tech. Rep.). Cambridge, MA: Man Machine Systems Laboratory, MIT.
- Steiger, J. H. (2007). Understanding the limitations of global fit assessment in structural equation modeling. *Personality and Individual Differences*, 42(5), 893–898.
- Sutcliffe, A. G., Poullis, C., Gregoriades, A., Katsouri, I., Tzanavari, A., & Herakleous, K. (2018, March). Reflecting on the Design Process for Virtual Reality Applications. *International Journal of Human-Computer Interaction*, 35(2), 168–179.

- 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., 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.
- Yang, X. J., Wickens, C. D., & Hölttä-Otto, K. (2016). How users adjust trust in automation: Contrast effect and hindsight bias. In *Proceedings of the Human Factors and Ergonomics Society* (Vol. 60, pp. 196–200).

## 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 a Ph.D. pre-candidate in the Department of Industrial & Operations Engineering at the University of Michigan. Her research interest includes trust in automation and design of decision aids on trust calibration. Prior to joining the University of Michigan, she completed a B.S. in Psychology in Zhejiang University, China in 2016.

**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.