The IMPACT of Agent Transparency on Human Performance

Kimberly Stowers, Nicholas Kasdaglis, Michael A. Rupp, Olivia B. Newton, Jessie Y. C. Chen, and Michael J. Barnes

***Abstract*—The primary purpose of this article is to determine the impact of a simulated agent's transparency on human performance and related variables, such as response time, workload, and trust calibration. The agent supports participants as they complete a base defense task by managing a team of heterogeneous unmanned vehicles and serves as a decision aid to the human. Three conditions of transparency are explored. In condition 1, the agent displays only the basic information (map of vehicle location and proposed routes). In condition 2, the agent displays the basic information and an explanation of its reasoning. In condition 3, the agent displays the basic information, reasoning, and uncertainties involved in the plans. Results show that participants exhibit better performance and trust calibration in the high-transparency conditions without perceiving a significant increase in workload. However, response time also increased, likely due to the additional processing time needed for conditions with more information. Overall, our findings indicate that the increased agent transparency can improve human-agent decision making and performance, but with a small cost of the efficiency (timeliness) of task completion.**

***Index Terms*—Human-agent team, human–machine system, performance, response time, trust, workload.**

I. Introduction

THE emergence of advanced technological capabilities has allowed humans to team with machines in a variety of contexts and with increasingly complex demands. In both civilian and military workplaces, it is increasingly common to see machines perceive, analyze, and act upon the world in roles of decision aids and assistants. These machines have become known as "agents" [1]. More recently, research has shifted to focus on creating collaborative operations in human-agent teams, where agents operate as teammates rather than tools [39]. In a military context, this may include the refinement of mission command to include information and planning systems that are capable of engaging in mixed-initiative decision making, or flexible collaboration that evolves with the task [2]. Efforts by the National Science Foundation and other entities suggest that such refinement is also needed in other domains (e.g., [3]).

To that end, the purpose of this article is to discuss the design, implementation, and results of a research study with a simulated agent that aided humans in completing a military command and control task by managing unmanned vehicles (UxVs). We specifically examined the impact of the agent on human performance, workload, response time, and trust. We devote the next section of the article to an exploration of the theoretical underpinnings of the research completed, then detail the research study, our findings, and their implications. Finally, we discuss steps to be taken for future research.

Manuscript received June 5, 2018; revised April 7, 2019 and January 24, 2020; accepted February 15, 2020. Date of publication April 16, 2020; date of current version May 18, 2020. This work was supported by the U.S. Department of Defense Autonomy Research Pilot Initiative, under the Intelligent Multi-UxV Planner With Adaptive Collaborative/Control Technologies (IMPACT) project. This article was recommended by Associate Editor G. Coppin. *(Corresponding author: Kimberly Stowers.)*

Kimberly Stowers is with the Department of Management, University of Alabama, Tuscaloosa, AL 35487 USA (e-mail: kim.stwrs@gmail.com).

Nicholas Kasdaglis is with the System Engineering Technical Center, MITRE Corporation, McLean, VA 22102-7539 USA (e-mail: nickkasdaglis@gmail.com).

Michael A. Rupp is with the Institute for Simulation and Training, University of Central Florida, Orlando, FL 32816 USA (e-mail: mrupp@knights.ucf.edu).

Olivia B. Newton is with the Department of Psychology, University of Central Florida, Orlando, FL 32816 USA (e-mail: obnewton@gmail.com).

Jessie Y. C. Chen is with the Human Research Engineering Directorate, U.S. Army Research Laboratory, Adelphi, MD 20783 USA (e-mail: jessie.chen@us.army.mil).

Michael J. Barnes is with the Human Research and Engineering Directorate, U.S. Army Research Laboratory, Ft. Huachuca, AZ 20783-1138 USA (e-mail: michael.j.barnes.civ@mail.mil).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/THMS.2020.2978041

II. Agent Transparency and Human Performance

Over the past 50 years, research has shown that a multitude of variables influences the human-agent team performance [4]. These variables continue to change and evolve with the state of the science in human–agent interaction. For example, advancements in machine automation and capability led some to realize that the *transparency* of an agent's conceptual processes may benefit such interaction by facilitating human acceptance, trust, and other relational factors (e.g., [5], [6]). Indeed, given the technological capability of machines today, transparency may not be just beneficial—but crucial—to successful human-agent interaction. Researchers across many fields have come to a shared consensus, with solutions to the complex problem of agent transparency growing to include new paradigms, such as situation awareness-based transparency (SAT; [7]) and explainable artificial intelligence (XAI; [8]).

A. Theoretical Approach

In human–agent interaction, the transparency may be defined in relation to information sharing or the clarity of conceptual and abstract constructs for human understanding [7]. Building from this, researchers have operationalized transparency as an agent

Fig. 1. SAT model, adapted from Chen *et al.* [6].

sharing information to support human understanding, including its contextual awareness and decision-making processes [10].

The relevance of transparency to human–agent interaction lies in its ability to improve the interaction itself. The benefit can be found not only in changing human attitudes in the interaction but also in improving the human-agent team's overall performance. Chen *et al.* [7] previously argued that the performance improvement in this way is tied to the agent's ability to facilitate SA in the human teammate—which would then lead to better decision making by the human (and better outcomes by the team). SA, or the perception of elements in a specific time and place [11], [12], can be parsed into three levels: perception of the current situation, comprehension of the situation, and projection as it relates to the perceiver and situation in the future [13].

To maximize SA in human–agent interaction, Chen *et al.* proposed the SA-based agent transparency (SAT) model (see Fig. 1, [6], [7]). Similar to Endsley's (1995) model of SA [13], SAT evolves according to the agent's communication of its perception (L1), comprehension (L2), and projection (L3) of future events. At the same time, SAT attempts to account for an agent's beliefs, desires, and intentions [14], as well as its purpose, process, and performance [15]. Furthermore, its communication of projection reaches beyond "what will happen next?" to the inclusion of uncertainty, a construct implicit to human decision making [16].

Human decision making in conditions of uncertainty includes the distinct feature that an action's consequences may be unknown, or at least not known with certainty [17]. More and more, such decisions are being made in the context of human–machine interaction, sometimes with a machine attempting to allay uncertainty by providing advice relevant to the decision-making process.

However, attempting to alleviate uncertainty in this way has some caveats, including a human overreliance on the machine when they trust or agree with its information [18]. In contrast, the implementation of the SAT model to shape the agent behavior can utilize uncertainty in a different way—specifically by making the human aware of the machine's own uncertainty or perceptions of uncertainty in the task environment. This may lead to a more accurate understanding of information that the human should consider, as well as encourage the human to maintain control in decision making [41].

B. SAT and the Calibration of Trust

Prevention of human overreliance on their machine counterparts has been a concern since automation was mainstreamed into society. With the increase of intelligence in machines (i.e., the ability to perceive and rationally act on the world [1], [42]), this concern has only grown. But how can appropriate reliance [18] in machines be fostered such that humans rely on the machine when it is helpful but not rely on it when it is wrong or otherwise unable to help? Some scholars suggest the key lies in *trust calibration*. For example, previous trust frameworks have suggested an operator's appropriate reliance on the agent can be called calibrated trust [19], [20] or appropriate learned trust [21]. However, any method to measure this phenomenon would be naturally complex—it must take into account a human's attitude toward the agent as well as the human's decision making itself. Yet decisions made by humans are not always as affected by agent teammates. Nevertheless, an exploration of trust calibration in relation to performance may offer insights into the human decision-making process in human-agent teams.

III. Prior Work

Multiple efforts have been pursued to determine the utility of the SAT model for improving human performance when interacting with assistive agents in an UxV management task. UxV management has evolved with a goal of allowing humans to supervise or control multiple UxVs [22]. However, controlling a team of multiple UxVs places additional physical and cognitive workload on the human, which may make it difficult to maintain SA, limiting mission success. In order for agents to alleviate this strain on humans, they must not only be able to track UxVs but also facilitate decision making in their human supervisors through effective information sharing and advising. To that end, a project funded by the Department of Defense's Autonomy Research Pilot Initiative (Intelligent Multi-UxV Planner with Adaptive Collaborative/Control Technologies—IMPACT; [23])

TABLE I
OPERATIONAL DEFINITIONS OF HUMAN DECISIONS (LABELED "HUMAN CHOICE" RELATING TO AGENT RECOMMENDATIONS (LABELED "AGENT CHOICE")

Correct Plan	Agent Choice	Human Choice	Usage	SDT Code
A	A	A	Proper Use	Hit
B	A	B	Correct Reject	Correct Reject
A	A	B	Disuse	Miss
B	A	A	Misuse	False Alarm

SDT = Signal detection theory.

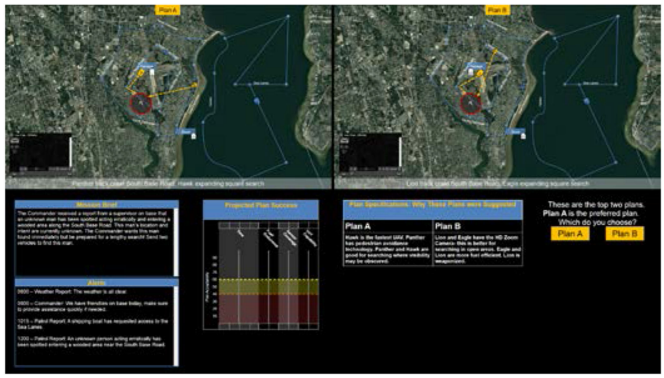

Fig. 2. Simulated agent interface for condition 1, containing Level 1 + 2 information [26].

explored several considerations for improving performance in UxV management, including information design in the agent.

As a part of this effort, [24] explored the utility of the SAT model for designing an agent that assists human decision making in a heterogeneous multi-UxV management task in which participants had to determine courses of actions or "plays" [25] for UxVs to pursue in order to meet mission goals. In this context, a "play" is a plan to coordinate specific UxVs to complete a task. For example, a human operator may choose to send three UxVs—one aerial, one surface, and one ground—to complete a patrol in the area of concern.

To support this effort, Mercado *et al.* [24] designed a simulated agent that assisted with decision making in a play-calling task by giving a recommendation and explaining its recommendation by sharing relevant information at the levels specified in the SAT model. Their primary goal was to measure whether the agent adding information according to the SAT model improved the human's decision-making performance.

To further contextualize the performance outcomes, Mercado *et al.* [24] additionally created a measure of *trust calibration* in which the signal detection theory (SDT) is applied to participants' decision making in the study. Like the traditional SDT, which measures a human's ability to differentiate between or identify stimuli depending on its salience [26], Mercado and his colleagues measured participants' sensitivity to the agent's accuracy in its decision recommendations. This evidence of participants' responses to the recommendations (signals) served as a measurement of trust calibration [27]. Specifically, Table I lists how codifying the simple act of choosing between two plans can give the partial evidence of trust calibration in this way. In particular, choices tracked in this way can be used to create a measure of *perceptual sensitivity* representing the human's choice relative to the signal (agent's recommendation). Through the examination of this phenomenon, Mercado and his colleagues found that the inclusion of projection information (Level 3) significantly increased perceptual sensitivity compared with the basic Level 1 information.

Through an experimental study with the simulated agent, Mercado and his colleagues established that, compared with a baseline agent that provided only Level 1 (L1) information, an explanation of the agent's reasoning (Level 2 or L2) led to improved performance on the task. An additional display of future projections (Level 3 or L3)—including the agent's uncertainties regarding those projections—led to an increase in the human's ability to determine when the agent's suggested plan was incorrect. However, because the implementation of Level 3 information conflated both projection and uncertainty, they were unable to delineate which improvements were due to the projection information versus the uncertainty information.

Results showed no main effects of transparency on response time or workload, though individual differences did play a role in response time. Many factors could have affected these outcomes, including the nature of the task, as well as the design of the agent interface. The question remains whether such findings are generalizable to other interfaces or domains.

IV. CURRENT STUDY

To determine whether the results by Mercado *et al.* [24] could be replicated with a new interface and specified at a more granular level with respect to Level 3 and the role of uncertainty, we chose to develop and test a simulated agent embedded in a similar UxV-management task. Specifically, we sought to understand if performance improves more with the addition of complete Level 3 information (projection and uncertainty) than with only partial Level 3 information (projection information on its own). Additionally, we sought to link general performance (correct and incorrect decisions) to trust calibration (appropriate agreement with agent) and self-reported trust. Finally, we sought to determine if Mercado and his colleagues' prior findings would generalize to the newly developed interface.

We developed three versions of the interface for the simulated agent—each differing by the amount of information being shared with the human. Using the SAT model as the guiding framework, the interface conditions included: L1+2 (agent state + reasoning), L1+2+3 (agent state + reasoning + projection without uncertainty), and L1+2+3+U (agent state + reasoning + projection + uncertainty). Figs. 2–4 show each interface, respectively. As can be seen, the interface grows increasingly complex with each condition, raising a concern that the processing

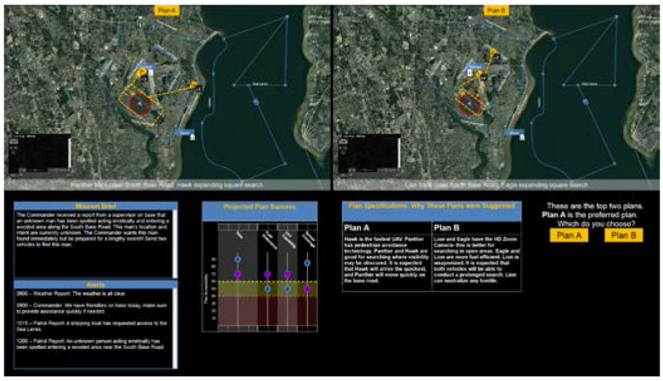

Fig. 3. Simulated agent interface for condition 2, containing Level 1 + 2 + 3 information [26].

Fig. 4. Simulated agent interface for condition 3, containing Level 1 + 2 + 3 + U information [26].

time and cognitive load may increase. A thorough discussion of the components of the interface, and results of usability-focused tests, has been published [28].

Transparency was manifested through alterations of three key tiles in the display, which participants were trained on how to read: plan maps, "projected plan success" graph, and "plan specifications" text. This allowed us to communicate the same information in multiple scalable ways, both graphically and in text. For example, uncertainty information was provided on the map by making vehicles, which the agent had uncertain information on less opaque. It was displayed in the "projected plan success" tile by making the circles representing each plan unfilled (meaning no color inside the circle). Finally, it was displayed in the "plan specifications" text as the orange text written separately from the primary text provided for reasoning.

For the purpose of this study, the simulated agent (via the interfaces above) rank-ordered the plans so that the plan with the highest rank was always recommended as "Plan A." The agent only gave the correct recommendation in five out of every eight scenarios. Thus, we were able to detect cases in which participants may have agreed with the agent when they should not, or vice versa, thus, allowing for the ability to analyze performance and trust calibration according to the process operationalized in Table I. Next, we discuss the implementation of the study and the resulting findings.

V. Method

A. Study Design

We designed and implemented a repeated measures study design with transparency being the only independent variable of interest. Transparency was manipulated at three levels: L1+2 (condition 1), L1+2+3 (condition 2), and L1+2+3+U (condition 3). Although all participants received all three conditions, the order in which they received those conditions was scheduled using a Latin Square design to avoid order effects.

B. Hypotheses

We sought to find the support for the following hypotheses, in line with Mercado *et al.* [24].

1) H1: There will be the main effect of transparency condition on performance.
2) H2: There will be no main effect of transparency condition on response time.
3) H3: There will be no main effect of transparency condition on self-reported workload.
4) H4: There will be a main effect of transparency condition on calibrated trust.
5) H5: There will be a main effect of transparency on self-reported trust, specifically when controlling for preinteraction bias toward machines.

C. Participants

We recruited students from a major university in the United States. Students were paid $15/h to participate. Data from a total of 53 participants between the ages of 18 and 36 ($M = 21.7$, $SD = 3.64$) were analyzed for this study. A total of 29 participants identified as men, 23 as women, and 1 as "other."

D. Materials

Study materials were integrated into a single testbed, which consisted of: 1) a simulated agent with the interface being manipulated according to the study design; and 2) a battery of surveys and tests to pursue our primary research questions of interest, as well as some exploratory questions. A full technical report on the study development, challenges, primary analyses, and exploratory analyses will be available on the Defense Technical Information Center (see [29]).

1) Simulated Agent Interface: As discussed above, we utilized a simulated agent modeled after Air Force Research Laboratory's Fusion testbed [43] for this experiment (see Fig. 2 for our version). The interface was altered for each condition of the experiment.

2) Surveys and Tests: The primary variables tracked in this study included performance, response time, workload, and trust. Performance was measured by the participant selection of the correct (i.e., more optimal) or incorrect (less optimal) play offered by the simulated agent. Response time was analyzed in milliseconds and reflected the time it took to select and submit a play for the experimental scenarios. Self-reported workload was

measured via the NASA-TLX questionnaire, which is reported to have a test-retest correlation of 0.83 [30].

Trust was measured objectively as *trust calibration* through the adaptation of the SDT to the participants' choices. This adaptation is modeled off of methods used in prior work by Mercado *et al.* [24]. Specifically, a measure of perceptual sensitivity, d', was created to represent the appropriate use of the agent's decision recommendation, according to Table I.

Trust was also measured subjectively using a part of an integrated survey developed by Mercado *et al.* [24]. Specifically, the survey featured a combination of new items as well as existing items from a previously validated trust measures [31] and was broken down according to the four dimensions of automated functions [32]. These dimensions include information acquisition, information analysis, decision and action selection, and action implementation. Although we collected information across the same four dimensions as Mercado and his colleagues, we focused on—and analyzed—the findings related to only two dimensions: integrating/displaying information and suggesting decisions. We determined these dimensions ad hoc (after completing the display design but before commencing the study) because of their relevance to the display we created, which was focused on combining and displaying information as a part of decision suggestion.

To account for extraneous variation in trust toward the agent, internal biases that might impact the development of trust were recorded via an implicit association test [33]. The Implicit Association Test was originally developed to operationalize humans' automatic association between constructs and stimuli (for example, crime and race/ethnicity) and has been adapted for use in a wide variety of domains. We used a version formulated to detect biases against machines. For example, our version had participants react to words such as "machine" and "automation" in comparison with words such as "human" and "person."

Performance and response time were recorded automatically in the study testbed, while the Implicit Association Test, trust, and workload surveys were each given via the DUJO skill assessment software [34] before training and after completion of the study. Trust calibration was calculated based on the automatically recorded decisions throughout the study.

E. Procedure

Participants received up to 45 min of training through a self-paced, narrated PowerPoint presentation, which included embedded knowledge checks. The training familiarized participants with the task they were to complete, as well as the interfaces they would interact with to complete the task. The purpose of completing the training was to minimize the effects of a learning curve on the study design and experimental conditions. Thus, the training concluded with eight evaluation missions to test participant's ability to apply the knowledge to the task. Participants completed knowledge checks throughout the training presentation and prior to the evaluation missions. Participants who were not able to provide correct responses were given the remedial/repeated presentation of the materials they misunderstood. If they were still unable to correctly respond to knowledge checks after two attempts, they were paid and dismissed. Upon the successful completion of training, participants were subjected to the experimental manipulation in which they were randomly assigned to complete three blocks of eight scenarios or missions (where each block was a different condition of the interface). Within each condition, the agent suggested correct courses of action five times, and incorrect courses three times.

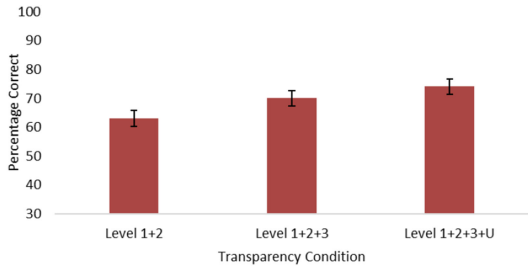

Fig. 5. Mean percentage of correct responses per condition. Error bars represent standard error of the mean.

VI. Analysis and Results

The data were analyzed using IBM SPSS version 24 and Microsoft Excel. Multiple repeated measures analysis of variance (ANOVA) and multivariate ANOVA (MANOVA) were completed to detect differences between conditions. A Bonferroni correction was applied to all pairwise comparisons explored in the analyses. Means and standard deviations are reported to demonstrate the degree of difference between conditions, where significant.

A. Performance

We first completed a repeated measures ANOVA on the percentage of correct responses given per condition. Results showed a significant main effect of transparency, $F(2, 104) = 10.31$, $p < 0.001$, $\eta^2 = 0.096$. Pairwise comparisons showed a significant difference ($p < 0.05$) between conditions 1 ($M = 0.62$, $SD = 0.2$) and 2 ($M = 0.70$, $SD = 0.14$). A significant difference ($p < 0.001$) was also found between conditions 1 and 3 ($M = 0.76$, $SD = 0.17$; and see Fig. 5).

In order to delineate performance specific to participants' ability to detect correct recommendations (hits) or detect incorrect recommendations (correct rejections), we completed a repeated measures MANOVA on hits and correct rejections. Results showed a significant main multivariate effect of transparency, *Wilk's Lambda* $= 0.80$, $F(4, 206) = 6.05$, $p < 0.001$. There was a significant main univariate effect of transparency on hits, $F(2, 104) = 10.663$, $p < 0.001$, $\eta^2 = 0.093$, but not on correct rejections. Pairwise comparisons (see Fig. 6) for hits indicated a significant difference ($p < 0.01$) between conditions 1 ($M = 0.59$, $SD = 0.23$) and 2 ($M = 0.72$, $SD = 0.21$), as well as a significant difference ($p < 0.001$) between conditions 1 and 3 ($M = 0.75$, $SD = 0.19$). Pairwise comparisons for correct rejections showed no significant differences. This suggests increased SAT improved participants' abilities to identify when the agent

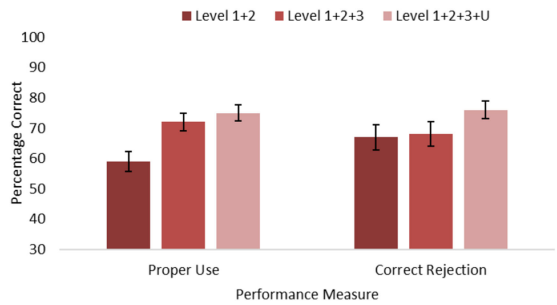

Fig. 6. Percentage correct for both proper use and correct rejection rates for all three transparency levels. Error bars represent standard error of the mean.

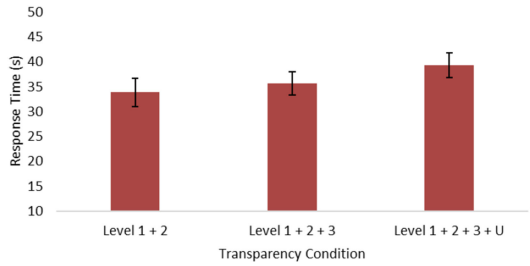

Fig. 7. Response time for all three transparency levels. Lower numbers indicate better response time; error bars represent standard error of the mean.

was making a correct recommendation, but not necessarily identify when the agent was making an incorrect recommendation.

B. Response Time

We completed a repeated measures ANOVA on the response time per condition. Results showed a significant main effect of transparency, $F(2, 104) = 4.37$, $p = 0.015$, $\eta^2 = 0.02$. Pairwise comparisons showed a significant difference ($p < 0.05$) between conditions 1 ($M = 33.65$ s, $SD = 19.68$ s) and 3 ($M = 39.47$ s, $SD = 16.6$ s; and see Fig. 7). Although the addition of both basic projection and uncertainty information resulted in a significantly increased response time for participants, this difference in time was a total of about 6 s. A much smaller increase in time was noted for the addition of basic projection information alone (condition 2, $M = 35.68$ s, $SD = 16.23$ s). The increase in time between conditions 2 and 3 was similarly minute in nature.

C. Self-Reported Workload

We first completed a repeated measures ANOVA on the total NASA-TLX score for each block. Results showed no significant effect of transparency on workload, $F(2,104) = 0.858$, $p = 0.427$, and $\eta^2 = 0.003$, suggesting that the additional information did not increase participant workload during decision making.

As an exploratory measure to gain a greater understanding of transparency's effect on the individual indices of workload, we completed a repeated measures MANOVA on the 6 NASA-TLX subscales, mental demand, physical demand, temporal demand, performance, effort, and frustration. Sphericity was violated for the "frustration" subscale resulting in the use of the Huynh–Feldt correction on this subscale. Test of within-subject effects indicated no significant multivariate effect and no significant univariate effects.

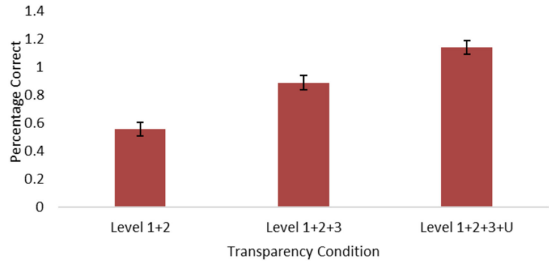

Fig. 8. Calibrated trust as calculated using d′ across conditions. Error bars represent standard error of the mean.

D. Trust Calibration

We completed a repeated measures ANOVA on our calculation of d' according to the methods detailed above. Results from a Kolmogorov-Smirnov test indicated a violation of normality. However, the visual inspection of the P–P plot suggested the data were acceptable, so we proceeded with the analysis. Results showed a significant main effect of transparency, $F(2, 104) = 9.018$, $p < 0.001$, $\eta^2 = 0.095$. Pairwise comparisons (see Fig. 8) showed a significant difference ($p < 0.01$) between conditions 1 and 3 ($M = 1.14$, $SD = 0.69$). However, there was no significant difference between conditions 1 and 2, or between conditions 2 and 3.

E. Self-Reported Trust

In addition to exploring the objective measure of trust calibration, we set out to determine the effect of transparency on self-reported trust—specifically as it relates to biases relevant to the interaction. To do this, we completed a repeated measures multivariate analysis of covariance on participants' responses to the integrating/displaying and suggesting actions indices of our trust survey, while controlling for responses to the Implicit Association Test.

Results showed a significant main multivariate effect of transparency, *Wilk's Lambda* $= 0.80$, $F(2, 48) = 2.63$, $p < 0.05$. There was no interaction effect between the transparency and Implicit Association Test. Tests of within subject effects indicated a significant main effect of transparency on trust in integrating/displaying information, $F(2, 102) = 3.48$, $p < 0.04$, $\eta^2 = 0.02$. Pairwise comparisons were completed and no significant differences were found among pairs (see Fig. 9).

Tests of within subject effects also indicated a significant main effect of transparency on trust in suggesting decisions, $F(2, 102) = 4$, $p < 0.03$, $\eta^2 = 0.02$. Pairwise comparisons were completed and a significant difference ($p < 0.05$) between conditions 1 ($M = 71.96$, $SD = 14.3$) and 2 ($M = 75.52$, $SD = 13.59$) was detected. The addition of uncertainty information (condition 3) resulted in a higher level of self-reported trust in suggesting decisions than condition 1, but a lower level of self-reported trust in suggesting decisions than condition 2 (see Fig. 9). However, there was no significant difference or notable

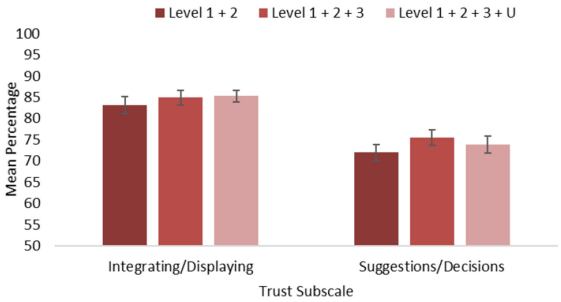

Fig. 9. Trust in integrating and displaying information, as well as suggesting decisions, for all three transparency levels. Error bars represent standard error of the mean.

effect size separating condition 3 from the other two conditions, meaning that the addition of uncertainty information on its own is not statistically supported for impacting self-reported trust.

VII. Discussion

The aim of this study was to examine how a human operator, possessing ultimate decision authority, will perform in a mixed-initiative task [35] in which s/he is aided by an agent. Fundamental to this examination was the acknowledgment that: 1) the agent may perform less than perfectly depending on its environment and context; and 2) the human operator may not understand and account for the agent's inherent limitations in its perception, reasoning, and projection. Operationally, the combination of these two presuppositions has been found to result in disuse or overreliance in the automation [18]. Therefore, to avoid these undesirable effects, human reliance on the agent must be appropriately calibrated to the capabilities of the agent.

Prior research has suggested that the demonstration of agent transparency or observability [20] can provide the means for appropriate trust and reliance in automation. This, in turn, can allay one of the most common challenges to human interactions with agents—specifically agent's capability of communicating to the human and its situation perception, reasoning, and projections [36], as proposed in the SAT model [7]. As discussed in Section IV, the study detailed, herein, evaluated the three conditions of transparency according to this model. That is, the agent communicating information at the following levels: L1+2 (agent state + reasoning); L1+2+3 (agent state + reasoning + projection without uncertainty); L1+2+3+U (agent state + reasoning + projection + uncertainty).

The results of this article uphold prior findings by Mercado *et al.* [24] showing that both human performance and trust calibration improve with the inclusion of greater transparency and that no detriment to workload is imposed. However, a key difference appeared in the impact of transparency on response time. Whereas Mercado and his colleagues found no effect of transparency on the response time, we found that participant's response time increased when exposed to uncertainty information. This difference in finding may be due to the inclusion of uncertainty in its own condition, or due to differences in the interface design itself. Uncertainty information was not linked to the reliability of the agent (meaning reliability was consistent across displays of information with/without uncertainty), so any design-related impacts are likely caused by characteristics, such as the use of text information or the sheer amount of information being presented in conditions with increased transparency.

A. Key Takeaways

Concerning the present study independently, the evaluation of each level of transparency described in our study could suggest the application and adoption of transparency enabling displays that share SAT information for human-agent teaming. Moreover, utilizing the construct of the SAT model, these principles can be applied systematically through the explicit representation of the following:

1) perceptual field of interest;
2) weighting and evaluation of the multiobjective space; and
3) comparison of projected outcomes (including those with high contextual uncertainty).

These attributes or challenges map to the specific SAT Levels 1–3 and the conditions that were evaluated in this study. Broadly, the purpose of this study was to explore which levels of transparency enabled appropriate calibration of trust, performance, and workload.

The first criterion is that adding to the level of transparency should improve human trust calibration relating to the agent's recommendations [24]. Results, herein, showed that the trust calibration was increased from condition 1 to 2 and 1 to 3, but not 2 to 3. Specifically, when the agent revealed its vulnerabilities in reasoning, projection, and uncertainty, operators demonstrated higher sensitivity and better decision making, accepting the agent's recommendation when it was correct, and correctly rejecting the agent's recommendation when it was wrong. This improvement in decision making suggests that operators receiving adequate transparency information are more likely to engage in appropriate reliance [18]. Interestingly, we found that the self-reported trust in the agent's ability to suggest decisions did not increase continuously with transparency but tapered off in the most transparent condition (condition 3, which included uncertainty information). Trust in the agent was highest in the L1+2+3 condition (without uncertainty). However, it is unclear how this finding relates to trust calibration.

The second criterion is that such trust calibration should result in improved human-agent collaboration and problem solving. In other words, additional transparency should lead to the improvement of the overall performance of the human in the completion of tasks (e.g., correctness in choosing an action). As reported, we found that an increase in the level of transparency from SAT Level 1 to Level 2 and Level 1 to Level 3 resulted in a significantly higher rate of correct responses. Yet performance is multidimensional—especially in a mixed-initiative task such as UxV management—in this domain, decisions must be made accurately and in a timely fashion. It is not surprising, then, that our results showed an increase in decision-making time in the most transparent condition 3 of our evaluation—on average, an increase of 6 s. Yet no significant increase in decision-making time was found from condition 1 to 2. Thus, it is possible that

the effect of increased transparency on the timeliness of decision making is largely due to either the amount of information presented or the type of information presented. In our case, condition 3 represented both an increase in information presented and an addition of a new type of information (uncertainty). Future research can determine which of these is more likely to affect the decision-making time in mixed-initiative tasks.

The final criterion is that an increase in transparency should not be so overbearing as to increase workload. As one can infer from our reported increase in timeliness to decide, transparency may come at a cost. However, there was no significant self-reported increase in workload as transparency level was increased. This suggests that the quality of information presented according to the SAT model actually aids in human decision making and cognitive processing to a degree that does not further tasks the human beyond the workload already imposed by the task.

B. Limitations

Despite the significant findings indicated in our study, our effect sizes remained small. To understand why, it is once again helpful to consider prior findings by Mercado and his colleagues, which showed that transparency gives a remarkable increase in performance over baseline (in which baseline contains only L1 information). In this article, we did not include this baseline condition, and instead looked at only the effects of information from L1+2 onward. As can be seen in the results, the differences between conditions grew smaller with each additional level of transparency, with the final condition, L1+2+3+U, offering the smallest change in human performance and related variables.

Therefore, it is important to consider our findings as they both might relate to baseline transparency (L1) as well as control (no transparency). For example, how much of an effect would be detected between no information and L1 information (and so on)?

C. Future Work

Machines have transcended elementary purposes of performing repetitive processes and repeatable actions for increased efficiency and reliability. Today's emergent use of agents serves to offer cognitive aid to humans who are restricted to the boundaries of their cognition and computational capabilities. Such challenging contexts are characterized by problem spaces with the high dimensionality of elements and their interactions. These contexts are often temporally constrained and have weighty aspects of risk (e.g., loss of human life). Thus, in future complex dynamic environments, human operators will increasingly rely on agents to help in decision making and the execution of collaborative action.

Therefore, designers of today's agents must reject technology-centered innovations that parse and encapsulate human functions from those of the machine. Humans aided by agents to perform increasingly cognitive work require an embrace of the perspective that what is being created is a new joint cognitive system [37], [38]. Such a perspective treats agents as teammates [39]. In such a system, humans may develop attitudes (e.g., trust) and behaviors (e.g., interaction) that is appropriate to the capabilities of the agent, resulting in a truly flexible collaboration [2].

In particular, future work should be completed to further understand the role of information sharing between humans and their agent teammates in improving human-agent performance. This includes first the delineation of the quantity of information with the type of information. Next, we must apply our knowledge of SAT and information sharing to a two-way human-agent teaming task, wherein the agent is given the opportunity to learn from the human [40]. Finally, we must examine the applicability of SAT to domains not yet studied.

Acknowledgment

The authors would like to thank D. Barber, J. Harris, and R. Wohleber for their assistance in testbed development and technology management and also G. Calhoun and M. Draper for their input during project development.

The author's affiliation with The MITRE Corporation is provided for identification purposes only and is not intended to convey or imply MITRE's concurrence with, or support for, the positions, opinions, or viewpoints expressed by the author. This article has been approved for Public Release; Distribution Unlimited; Case Number 20-0054.

References

[1] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*. Upper Saddle River, NJ, USA: Prentice Hall, 2009.

[2] G. Tecuci *et al.*, "A tool for training and assistance in emergency response planning," in *Proc. 40th Annu. Hawaii Int. Conf. Syst. Sci.*, Waikoloa, HI, USA, Jan. 3–6, 2007.

[3] NSF, "10 Big Ideas for Future NSF Investments," May 2016. [Online]. Available: www.nsf.gov/about/congress/reports/nsf_big_ideas.pdf

[4] K. Stowers, J. Oglesby, S. C. Sonesh, K. Leyva, C. Iwig, and E. Salas, "A framework to guide the assessment of human–machine systems," *Human Factors*, vol. 59, no. 2, pp. 172–188, Mar. 2017.

[5] J. L. Herlocker, J. A. Konstan, and J. Riedl, "Explaining collaborative filtering recommendations," in *Proc. ACM Conf. Comput. Supported Cooperative Work*, Philadelphia, PA, USA, 2000, pp. 241–250.

[6] A. L. Thomaz and C. Breazeal, "Transparency and socially guided machine learning," in *Proc. 5th Int. Conf. Develop. Learn.,* Bloomington, IN, USA, 2006.

[7] J. Y. C. Chen, K. Procci, M. Boyce, J. L. Wright, A. Garcia, and M. Barnes, "Situation awareness-based agent transparency," Army Res. Lab., Adelphi, MD, USA, Tech. Rep. ARL-TR-6905, Apr. 2014.

[8] D. Gunning, "Explainable artificial intelligence (XAI)," in *Proc. IJCAI Workshop Deep Learn. Artif. Intell.,* New York, NY, USA, Jul. 10, 2016.

[9] Princeton University, "Transparency," 2010. [Online]. Available: http://wordnetweb.princeton.edu/perl/webwn?s=transparency

[10] J. B. Lyons, "Being transparent about transparency: A model for human-robot interaction," in *Trust and Autonomous Systems: Papers From the 2013 AAAI Spring Symposium*, D. Sofge, G. J. Kruijff, W. F. Lawless, Eds. Menlo Park, CA, USA: AAAI Press, 2013, pp. 189–204.

[11] M. R. Endsley and D. G. Jones, *Designing for Situation Awareness: An Approach to User-Centered Design*. Boca Raton, FL, USA: CRC Press, 2012.

[12] N. A. Stanton, P. R. G. Chambers, and J. Piggott, "Situational awareness and safety," *Saf. Sci.*, vol. 39, pp. 189–204, Dec. 2001.

[13] M. R. Endsley, "Toward a theory of situation awareness in dynamic systems," *Human Factors*, vol. 37, no. 1, pp. 32–64, Mar. 1995.

[14] A. S. Rao and M. P. Georgeff, "BDI agents: From theory to practice," in *Proc. 1st Int. Conf. Multiagent Syst.*, 1995, vol. 95, pp. 312–319.

[15] J. D. Lee, "Trust, trustworthiness, and trustability," in *Proc. Workshop Human Mach. Trust Robust Auton. Syst.*, Ocala, FL, USA, Jan./Feb. 2012.

[16] A. Tversky and D. Kahneman, "Judgment under uncertainty: Heuristics and biases," *Science*, vol. 185, no. 4157, pp. 1124–1131, Sep. 1974.

[17] R. W. Proctor and T. Van Zandt, *Human Factors in Simple and Complex Systems*. Boca Raton, FL, USA: CRC Press, 2008.
[18] R. Parasuraman and V. Riley, "Humans and automation: Use, misuse, disuse, abuse," *Human Factors*, vol. 39, no. 2, pp. 230–253, Jun. 1997.
[19] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. C. Chen, E. de Visser, and R. Parasuraman, "A meta-analysis of factors impacting trust in human-robot interaction," *Human Factors*, vol. 53, no. 5, pp. 517–527, Oct. 2011.
[20] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human Factors*, vol. 46, pp. 50–80, 2004.
[21] K. Hoff and M. Bashir, "Trust in automation: Integrating empirical evidence on factors that influence trust," *Human Factors*, vol. 57, no. 3, pp. 407–434, May 2015.
[22] M. L. Cummings, A. Clare, and C. Hart, "The role of human-automation consensus in multiple unmanned vehicle scheduling," *Human Factors*, vol. 52, no. 1, pp. 17–27, Feb. 2010.
[23] M. Draper *et al.*, "Realizing autonomy via intelligent adaptive hybrid control: Adaptable autonomy for achieving UxV RSTA team decision superiority (also known as intelligent multi-UxV planner with adaptive collaborative/control technologies (IMPACT))," Air Force Res. Lab., Wright-Patterson, OH, USA, Air Force Research Laboratory-RH-WP-TR-2018-0005, 2018.
[24] J. E. Mercado, M. A. Rupp, J. Y. C. Chen, M. J. Barnes, D. Barber, and K. Procci, "Intelligent agent transparency in human–agent teaming for multi-UxV management," *Human Factors*, vol. 58, no. 3, pp. 401–415, Feb. 2016.
[25] C. A. Miller and R. Parasuraman, "Designing for flexible interaction between humans and automation: Delegation interfaces for supervisory control," *Human Factors*, vol. 49, no. 1, pp. 57–75, Feb. 2007.
[26] H. Stanislaw and N. Todorov, "Calculation of signal detection theory measures," *Behav. Res. Methods, Instrum. Comput.*, vol. 31, no. 1, pp. 137–149, Mar. 1999.
[27] J. Mercado, M. A. Rupp, J. Y. C. Chen, D. Barber, K. Procci, and M. Barnes, "Effects of agent transparency on multi-robot management effectiveness," Army Res. Lab., Adelphi, MD, USA, Tech. Rep. ARL-TR-7466, Sep. 2015.
[28] K. Stowers, N. Kasdaglis, O. Newton, S. Lakhmani, R. Wohleber, and J. Chen, "Intelligent agent transparency: The design and evaluation of an interface to facilitate human and intelligent agent collaboration," in *Proc. Human Factors Ergonom. Soc. Annu. Meeting*, Washington, DC, USA, Sep. 19–23, 2016, pp. 1706–1710.
[29] K. Stowers, N. Kasdaglis, O. Newton, M. Rupp, M. Barnes, and J. Y. C. Chen, "Effects of situation awareness-based agent transparency information on human agent teaming for multi-UxV management," Army Res. Lab., Adelphi, MD, USA, to be published.
[30] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research," in *Advances in Psychology*, vol. 52, P. Hancock and N. Meshkati, Eds. Oxford, U.K., 1988, pp. 139–183.
[31] J. Jian, A. M. Bisantz, and C. G. Drury, "Foundations for an empirically determined scale of trust in automated systems," *Int. J. Cogn. Ergonom.*, vol. 4, no. 1, pp. 53–71, Jun. 2010.
[32] R. Parasuraman, T. B. Sheridan, and C. D. Wickens, "A model for types and levels of human interaction with automation," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 30, no. 3, pp. 286–297, Jun. 2000.
[33] A. G. Greenwald, D. E. McGhee, and J. L. Schwartz, "Measuring individual differences in implicit cognition: The implicit association test," *J. Personality Social Psychol.*, vol. 74, no. 6, pp. 1464–1480, 1998.
[34] DUJO-Mind, Sourceforge, Jul. 2013. [Online]. Available: Retrieved from: https://sourceforge.net/projects/dujomind/
[35] M. A. Goodrich, "On maximizing fan-out: Towards controlling multiple unmanned vehicles," in *Human–Robot Interactions in Future Military Operations*, M. G. Barnes and F. Jentsch, Eds. Surrey, U.K.: Ashgate, 2010, pp. 375–395.
[36] N. B. Sarter and D. D. Woods, "How in the world did we ever get into that mode? Mode error and awareness in supervisory control," *Human Factors*, vol. 37, no. 1, pp. 5–19, Mar. 1995.
[37] D. D. Woods and E. Hollnagel, *Joint Cognitive Systems: Patterns in Cognitive Systems Engineering*. Boca Raton, FL, USA: CRC Press, 2006.
[38] K. Christoffersen and D. D. Woods, "How to make automated systems team players," in *Advances in Human Performance and Cognitive Engineering Research*, vol. 2, E. Salas, Ed. Bingley, U.K.: Emerald Group, 2002, pp. 1–12.
[39] E. Phillips, S. Ososky, J. Grove, and F. Jentsch, "From tools to teammates: Toward the development of appropriate mental models for intelligent robots," in *Proc. Human Factors Ergonom. Soc. Annu. Meeting,* Las Vegas, NV, USA, Sep. 19–23, 2011, pp. 1491–1495.
[40] J. Y. C. Chen, S. G. Lakhmani, K. Stowers, A. R. Selkowitz, J. L. Wright, and M. Barnes, "Situation awareness-based agent transparency and human-autonomy teaming effectiveness," *Theor. Issues Ergonom. Sci.*, vol. 19, no. 3, pp. 259–282, Feb. 2018.
[41] H. Griethe and H. Schumann, "Visualizing uncertainty for improved decision making," in *Proc. 4th Int. Conf. Bus. Inform. Res. BIR,* Jan. 2005, pp. 1–11.
[42] M. Wooldridge and N. R. Jennings, "Intelligent agents: Theory and practice," *Knowl. Eng. Rev.*, vol. 10, no. 2, pp. 115–152, 1995.
[43] A. J. Rowe, S. E. Spriggs, and D. J. Hooper, "Fusion: A framework for human interaction with flexible-adaptive automation across multiple unmanned systems," in *Proc. Int. Symp. Aviation Psychol.*, 2015, pp. 464–469.