Intelligent Agent Transparency in Human–Agent Teaming for Multi-UxV Management

Joseph E. Mercado, U.S. Army Research Laboratory, Orlando, Florida, **Michael A. Rupp**, University of Central Florida, Orlando, **Jessie Y. C. Chen**, **Michael J. Barnes**, U.S. Army Research Laboratory, Orlando, Florida, **Daniel Barber**, and **Katelyn Procci**, University of Central Florida, Orlando

Objective: We investigated the effects of level of agent transparency on operator performance, trust, and workload in a context of human–agent teaming for multirobot management.

Background: Participants played the role of a heterogeneous unmanned vehicle (UxV) operator and were instructed to complete various missions by giving orders to UxVs through a computer interface. An intelligent agent (IA) assisted the participant by recommending two plans—a top recommendation and a secondary recommendation—for every mission.

Method: A within-subjects design with three levels of agent transparency was employed in the present experiment. There were eight missions in each of three experimental blocks, grouped by level of transparency. During each experimental block, the IA was incorrect three out of eight times due to external information (e.g., commander's intent and intelligence). Operator performance, trust, workload, and usability data were collected.

Results: Results indicate that operator performance, trust, and perceived usability increased as a function of transparency level. Subjective and objective workload data indicate that participants' workload did not increase as a function of transparency. Furthermore, response time did not increase as a function of transparency.

Conclusion: Unlike previous research, which showed that increased transparency resulted in increased performance and trust calibration at the cost of greater workload and longer response time, our results support the benefits of transparency for performance effectiveness without additional costs.

Application: The current results will facilitate the implementation of IAs in military settings and will provide useful data to the design of heterogeneous UxV teams.

Keywords: intelligent agent transparency, human–agent teaming, multi-UxV management

Address correspondence to Joseph E. Mercado, U.S. Army Research Laboratory–Human Research and Engineering Directorate, 340 Hulse Rd., Pensacola, FL 32508, USA; e-mail: joseph.mercado@med.navy.mil.

HUMAN FACTORS
Vol. 58, No. 3, May 2016, pp. 401–415
DOI: 10.1177/0018720815621206

INTRODUCTION

Mission effectiveness for heterogeneous unmanned vehicle (UxV) teams relies on rapid identification and management of uncertainties that can disrupt the team's ability to complete complex operations safely. To improve mission effectiveness, many of today's operators utilize complex human–machine systems to supervise unmanned systems, though this interaction can overwhelm the operator due to its high rate of information flow (Chen, Barnes, & Harper-Sciarini, 2011; Paas & Van Merriënboer, 1994). In particular, future scenarios will require a single human to supervise multiple unmanned systems, which could easily surpass the operator's span of control depending on mission complexity (Cummings & Mitchell, 2008; Lewis, 2013). Given the negative impact of operator overload, intelligent agents (IAs) have been developed to unload the operator while improving overall human–agent team performance (Bradshaw et al., 2008; Hardin & Goodrich, 2009; Hwang et al., 2008; see Chen & Barnes, 2014, for a review). Whereas there are multiple definitions of *IA*, we chose the definition that most closely relates to the IA we are affording the human in the present experiment. *IA* refers to an entity that possesses the following characteristics: autonomy, observation of the environment, action upon an environment, and activity toward achieving certain goals (Russell & Norvig, 2009).

Multirobot management by humans working with IAs is not without issues (Chen & Barnes, 2014). Past research has demonstrated that human operators sometimes question the accuracy and effectiveness of the output produced by IAs due to operator difficulty with understanding IA rationale, leading to reduced use of the IA and subsequent loss of performance (Linegang et al., 2006). Researchers have suggested that to

support operator situation awareness (SA) of the IA in its tasking environment, the agent needs to be transparent about its reasoning process and projected outcomes (Chen et al., 2014; Lee & See, 2004). Agent transparency is the IA's ability to communicate information to the human operator in a clear and efficient manner, which allows the operator to develop an accurate mental model of the system and its behavior, leading to calibrated trust in the system (Chen et al., 2014; Lee & See, 2004).

The effect of increased agent transparency on operator workload depends on the amount of information provided and whether the information is required for the operator's task performance (Lyons & Havig, 2014). If implemented appropriately, these additional interface elements can reduce operator workload by helping the operator understand what the agent is trying to achieve and what the operator can expect to happen, so the operator does not have to make these connections himself or herself (Chen & Barnes, 2014). However, more information does not always equate to relevant and good information. If the increased information-processing requirements caused by the additional information shown to the operator increase workload, the display may be seen as less usable and will be trusted less.

Lee and See (2004) recommend that the system display its purpose, process, and performance (3Ps) as well as respective histories. To avoid overwhelming the operator, the presentation should be in a simplified form, for example, integrated graphical displays and simplified text (Cook & Smallman, 2008; Neyedli, Hollands, & Jamieson, 2011). In the present research we sought to examine the level of information necessary to create an effective "transparent interface" that does not induce operator overload, specifically addressing three issues: operator performance, operator trust in the IA, and operator workload. The following section explains the concept of agent transparency and a framework on which our experimental design was based.

Agent Transparency and SA

Humans interacting with highly automated systems encounter multiple challenges: understanding the current system state, comprehending reasons for its current behavior, and projecting its future behaviors (Sarter & Woods, 1995). As a result, transparency in automated systems has become an essential research area (Lee & See, 2004). Although there are multiple definitions of agent transparency (Chen et al., 2014; Helldin, 2014; Lyons & Havig, 2014), we use the definition proposed by Chen and colleagues (2014): "Agent transparency is the quality of an interface pertaining to its abilities to afford an operator's comprehension about an intelligent agent's intent, performance, future plans, and reasoning process" (p. 2). The goal of transparency is not to relay all of a system's capabilities, behaviors, and decision-making rationale to the human operator. Ideally, agents should relay clear and efficient information as succinctly as possible to the human operator, thus enabling the operator to maintain proper SA of the system in its tasking environment without becoming overloaded (Chen et al., 2014; Lee & See, 2004).

In an effort to determine the essential information that must be provided to the operator to support agent transparency, we leveraged the SA-based agent transparency (SAT; Chen et al., 2014) model. This model, based on Endsley's (1995) SA model, described the information that the IA should convey to the human in order for him or her to have proper SA of the agent in its tasking environment. According to Endsley (1995), SA is "a[n individual's] state of knowledge" of a dynamic environment (p. 36). In Endsley's model, SA has three levels: perception (Level 1), comprehension (Level 2), and projection (Level 3). Similarly, the first level of the SAT model provides the operator with basic information about the IA's current state, goals, intentions, and plan of action; the second level provides the operator with information about the IA's reasoning process behind the plan of action, including rationale, capabilities, limitations, and trade-offs between different options; and the third level provides the operator with information regarding predicted consequences and the likelihood of a plan's success or failure (Chen et al., 2014).

The SAT model implies that incorporating three levels of transparency should support the operator's SA of an IA in its tasking environment, specifically the IA's intent, reasoning, projected outcomes, and uncertainty. The opera-

tor's SA of the IA's intent, reasoning, projected outcomes, and uncertainty should improve the operator's subjective trust as well as trust calibration, which is proper reliance when the IA is correct and correct rejection when the IA is incorrect (Lee & See, 2004). An appropriate calibration of trust in an IA is especially vital in high-risk situations, such as military operations, where over- and undertrust can be disastrous (de Visser et al., 2012; Freedy, DeVisser, Weltman, & Coeyman, 2007; Groom & Nass, 2007; Lee & See, 2004; Parasuraman & Riley, 1997). Although informing the operator of IA uncertainty seems like a counterintuitive way to improve trust calibration, understanding the limitations of a system has been reported to improve trust calibration and subjective trust in the decision-making process (Dzindolet, Peterson, Pomranky, Pierce, & Beck, 2003; Lee & See, 2004; Wang, Jamieson, & Hollands, 2009).

Current Study

We simulated a heterogeneous multi-UxV planning task, in which participants performed the role of an operator whose job was to work with an intelligent planning agent. The objective was to decide on which courses of action should be carried out by the UxVs to ensure mission success based on the commander's intent, vehicle capability, and environmental constraints. We utilized the Wizard of Oz (Riek, 2012) technique to simulate both current and future capabilities of the Intelligent Multi-UxV Planner With Adaptive Collaborative/Control Technologies (IMPACT) system, currently developed under the U.S. Department of Defense Autonomy Research Pilot Initiative. The IMPACT system combines "flexible, goal-oriented play-calling and human-autonomy interaction with cooperative control algorithms that provide near-optimal task assignment and path planning solutions as well as adaptive/reactive capability" (Draper, 2013, p. 1). In this context, a "play" refers to a scripted plan for coordinating multiple UxVs, which is augmented by the IA that chooses the appropriate assets for the specific mission (Douglass, 2013; Miller & Parasuraman, 2007).

Our goal was to examine the level of information necessary to create an effective and transparent interface to support human–agent teaming for management of multiple UxVs. Whereas there are various types of agent transparency with different levels of details, we focused on the transparency that would support the operator's SA of the agent in its tasking environment (based on the SAT model). We hypothesized that a greater level of transparency in the interface should promote effective trust calibration, which, in turn, should support overall human–agent team performance. Workload is a valid concern when additional information is added to an interface, as it may affect both trust in the system and perceived usability. We hypothesized that operator workload would decrease with increased transparency level, because the operator would not have to speculate on the reasoning behind the IA's recommended plans and projected outcomes. However, we also note that increased workload is a valid concern when additional information is added to an interface, and increased workload may reduce perceived usability in the system (Bevan & Macleod, 1994). Finally, prior research suggests that individual differences factors, such as operator spatial ability, attentional control ability, video gaming experience, and working memory capacity (Ahmed et al., 2014), may affect human–agent teaming (see Chen & Barnes, 2014, for a review). The effects of these factors on operator performance were also evaluated.

METHOD

Participants

Participants for this experiment included both undergraduate and graduate students from the University of Central Florida. Thirty young adults (18 men, 12 women) between the ages of 18 and 29 (M = 21.2, SD = 2.3) were recruited using an online participant pool. Participants were required to have normal or corrected-to-normal vision (including not being color-blind). They were compensated $15 per hour for their participation, which lasted approximately 3 hr.

Apparatus

Simulator. A customized computer-based simulator, based on the U.S. Air Force Research Laboratory's FUSION interface (Spriggs, Boyer, & Bearden, 2014), was utilized in the present

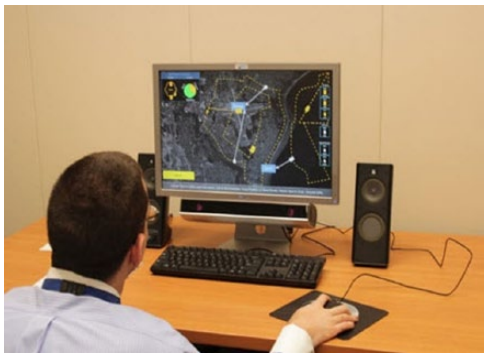

Figure 1. Desktop simulator system.

study (Figure 1). The FUSION interface itself was not utilized because it is not designed for experimental use. The simulator screen included several sections: a video window, where participants watched UxV movements and received intelligence messages (intel); a mission assignment window, where participants received a mission objective; and a decision window. In the decision window, participants received information on vehicle capabilities, mission synopsis, intel, and both of the IA's plan suggestions (Plan A and Plan B; Figure 2). After the mission assignment window, participants evaluated the two plans and selected the best plan based on their judgment. Participants were instructed to use three metrics to evaluate each plan (the same three metrics used by the IA): speed (how quickly each UxV can arrive or carry out the mission), coverage (how well the UxV can find a target based on its sensors), and capabilities (vehicles' appropriateness for the mission). Each UxV had a set of strengths and weaknesses (e.g., can travel long distances, is stealthy, or is weaponized) that could affect capabilities.

The system provided transparency to the human operator via a sprocket graphic, text box, and uncertainty information. The sprocket graphic provided information on the three important plan attributes: speed, coverage, and capabilities. The size of the wedge indicated the importance of the factor. The color of the wedge indicated how well the plan achieved this factor of the play, whereby green indicated high suitability and yellow indicated suboptimal suitability. The text box explained the environmental and UxV-related factors that affected the IA's decision, displaying information pertaining to each of the three plan attributes. For example, the text box would be used to explain why a specific UxV or route was used over another. In regard to UxV fit, large UxV icons represented UxVs that were better equipped for the play. For example, if speed was the most important factor, then the faster UxVs were larger.

Eye tracker. The SMI (SensoMotoric Instruments) Remote Eye-Tracking Device (RED) was used to collect ocular indices to measure visual attention (Hoffman & Subramaniam, 1995) and workload (Poole & Ball, 2006; Figure 1).

Surveys and tests. A short questionnaire captured basic demographic information and video game experience. Participants who reported playing action video games on either a daily or weekly basis were categorized as action video game players (AVGPs; n = 18; 17 men, one woman); all other participants were categorized as non-AVGPs (n = 12; 11 women, one man).

A nine-plate Ishihara color vision test was administered via PowerPoint. Normal color vision was required to understand the user interface fully. The Cube Comparison Test (Ekstrom, French, & Harman, 1976) and the Spatial Orientation Test (Gugerty & Brooks, 2004) were used to assess participants' spatial ability. Both tasks measured related but distinct components of spatial ability (Hegarty & Waller, 2004): spatial visualization (SpaV; the mental rotation of objects) and spatial orientation (SpaO; the reorientation of an environment). Using a median split, we classified 15 participants as high SpaV, low SpaV, high SpaO, or low SpaO, respectively. We used a version of the Operator Span task (Conway et al., 2005) to measure working memory capacity (WMC). Using a median split, we classified 17 participants as low WMC and 13 participants as high WMC.

Before beginning the experimental missions, participants were trained on the computer-based simulator. Training lasted approximately 1 hr using PowerPoint presentation and the simulator. The presentation introduced the procedures and protocols for participating in research experiment. Participants then completed training missions similar to those in the experimental missions for each transparency level. Feedback was given after each transparency level.

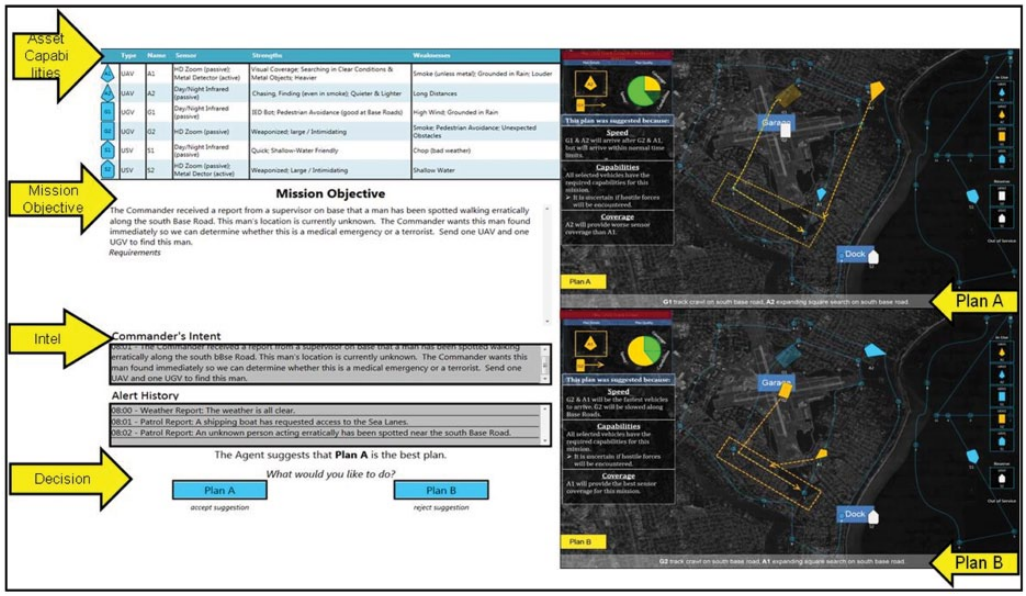

Figure 2. Simulator decision window.

After each training mission block, participants completed an in-house usability questionnaire with which they ranked how often they utilized each aspect on the decision window (e.g., sprocket graphic, text box).

After each mission block, participants completed the System Usability Scale, a 10-item summative usability survey (Brooke, 1996). Participants' perceived workload was evaluated with a computer-based version of the NASA Task Load Index questionnaire (NASA-TLX; Hart & Staveland, 1988). Finally, participants were asked to evaluate their trust in the IA along two dimensions, (a) information analysis (trust in the information and analysis displayed) and (b) decision and action selection (trust in the IA's suggestions and decisions; Parasuraman, Sheridan, & Wickens, 2000), using a modified version of the Trust Between People and Automation questionnaire developed by Jian, Bisantz, and Drury (2000).

Procedure

After being briefed on the purpose of the study and giving informed consent, participants completed the demographics questionnaire and the Ishihara test. Participants then received training on the tasks, which lasted approximately 45 min and consisted of PowerPoint slides and training missions performed using the simulator. During training, participants were informed that the IA was not always 100% accurate. After training, participants received 18 evaluation missions and were required to perform at least 12 missions correctly to move on to the experimental session. The evaluation lasted approximately 40 min. Participants were given a 5-min break, after which the eye tracker was calibrated.

The experimental session consisted of three counterbalanced blocks of eight missions. During each mission, participants controlled a team of UxVs (ground, air, and sea) in a military perimeter defense task. Participants received intel and the commander's intent and used them to choose one of the IA's recommended plans. The IA always recommended two plans: Plan A (the agent's top choice) and Plan B (the agent's secondary or backup choice). The IA's top choice was optimized for the most important metric, which was indicated by wedge size. For example, if the speed wedge was the largest, then the IA would optimize the play for speed to complete the mission as quickly as possible. The IA's backup choice was optimized for one of the other two metrics (coverage or capabilities if speed was the most important metric).

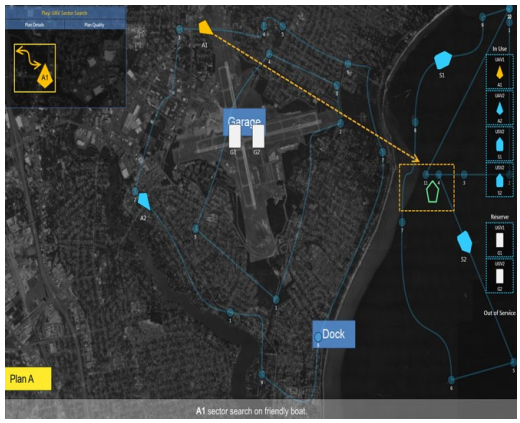

Figure 3. Transparency Level 1 simulator interface.

For three out of the eight missions, Plan B was actually the better plan, due to information that the IA was not aware of (e.g., intelligence, commander's intent). For example, the commander's intent could instruct the human to complete the mission as fast as possible, but because the IA was not aware of the commander's intent, it optimized for capabilities. The error rate was based on Wickens and Dixon's (2007) finding that a 70% reliability rate is the point at which unreliable automation was worse than a lack of automation in terms of performance.

The experiment used a within-subjects design with agent transparency as the independent variable with three levels (each block of experimental missions corresponded to one transparency level, which was counterbalanced by both order and missions). Transparency Level 1 consisted of a baseline condition, which provided basic plan information (which UxVs were used and the paths they utilized; Figure 3). Transparency Level 1+2 consisted of all the information provided by Level 1 plus the agent's reasoning and rationale behind recommending the plans via a text box and sprocket graphic (Figure 4). Transparency Level 1+2+3 consisted of all the information provided by the previous transparency levels plus projection of uncertainty information—opacity of the vehicle icons, road colors, sprocket graphic wedges, and bullet points in the text box indicated uncertainty as to whether the factor in question would contribute to a successful action (Figure 5). Participants were not shown probabilities or likelihood comparisons, just that the information was uncertain. For example, a speed wedge could be green (meaning this metric was well satisfied by this plan) but not completely opaque in color (meaning it was also uncertain). The specific reason for the uncertainty was listed in the text box. For example, environmental constraints (e.g., possible winds) may slow down certain vehicles, reducing speed.

Participants had 2 min to complete each mission. The experimental session lasted approximately 90 min. After each block, participants completed the NASA-TLX, the trust survey, and the System Usability Scale. In total, the experiment lasted approximately 4 hr.

Dependent Measures

Dependent measures included operator performance measures (correct IA usage, correct IA, and response time), participants' self-assessed trust in the IA, and workload measures (NASA-TLX and eye movement measures).

RESULTS

We completed a series of mixed analyses of variance (ANOVAs) and multivariate analyses of variance (MANOVAs) across all dependent measures (including individual differences data). All post hoc comparisons utilized a Bonferroni alpha correction. We report effect sizes in terms of η^2 instead of partial η^2 as these can be more easily compared across studies (Levine & Hullett, 2002).

Operator Performance

We evaluated operator performance by examining correct IA usage and correct IA rejections. A repeated-measures MANOVA on both measures across each transparency level was used to reduce pairwise error rate, since both were moderately correlated (rs = .26–.73) but not so strongly correlated that they warranted creating a composite measure. This analysis revealed a significant multivariate effect for transparency level using Wilks' lambda criterion, $F(4, 21) = 7.15$, $p = .001$, $\eta^2 = .58$, $\lambda = .42$.

Correct IA Usage Rates

Results for correct IA usage revealed a significant main effect of transparency level, $F(2,$

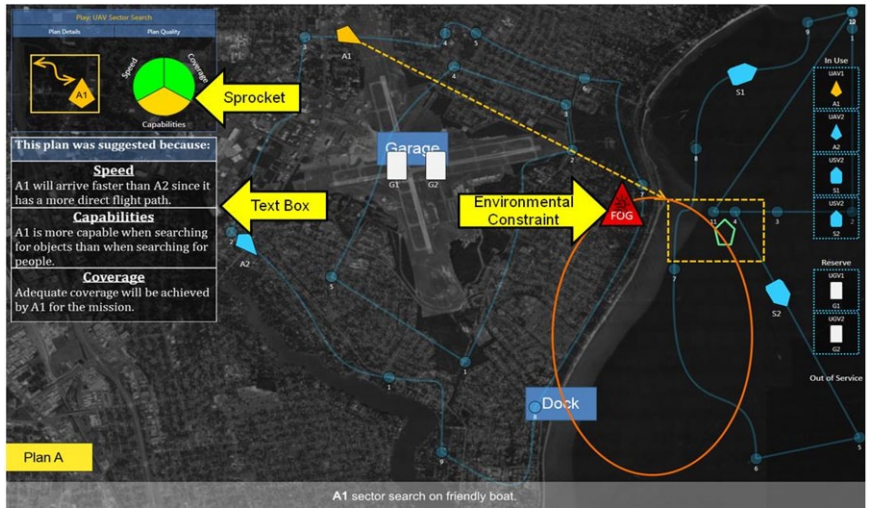

Figure 4. Transparency Level 1+2 interface. Includes interface items that describe agent rationale for recommended plan.

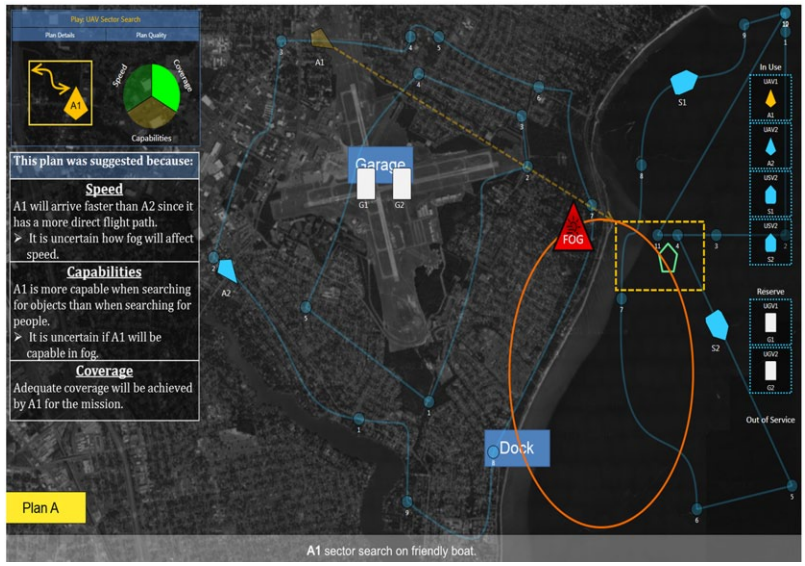

Figure 5. Transparency Level 1+2+3 interface. Transparent items suggest uncertainty of an asset, path, or sprocket graphic.

58) = 12.33, $p < .001$, $\eta^2 = .30$ (Figure 6; Table 1). Post hoc comparisons indicated participants' correct IA usage rates were significantly greater in Level 1+2+3 ($p < .001$) and Level 1+2 ($p = .003$) compared to Level 1.

Correct IA Rejection Rates

Results for correct IA rejection rates revealed a significant main effect of transparency level, $F(2, 58) = 15.03$, $p < .001$, $\eta^2 = .34$ (Figure 6; Table 1). Post hoc comparisons indicated that participants' correct IA rejection rates were greatest in Level 1+2+3, significantly greater than Level 1+2 ($p = .04$), which, in turn, is greater than Level 1 ($p = .013$).

There was a marginally significant interaction effect for working memory capacity, $F(2, 56) = 3.07$, $p = .054$, $\eta^2 = .01$. High-WMC

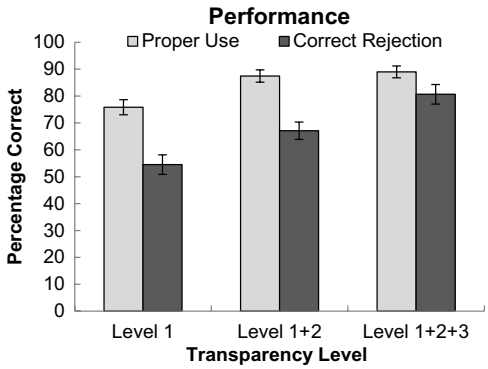

Figure 6. Percentage correct for both intelligent agent (IA) correct usage and IA correct rejection rates for all three transparency levels. Greater numbers indicate better performance. Error bars are standard error of the mean.

individuals outperformed low-WMC individuals in Level 1 ($d = 0.92$) but not in the other conditions (Figure 7).

Response Time

There was no significant difference in response time among the transparency levels (Table 1), nor were there any significant individual differences except for gaming experience. Results revealed a significant interaction effect for gaming experience, $F(2, 56) = 5.74$, $p = .005$, $\eta^2 = .17$ (Figure 8). AVGPs had quicker response times than non-AVGPs in both Level 1 ($d = 0.61$) and Level 1+2+3 ($d = 0.26$). AVGPs had quicker response time than non-AVGPs for Level 1+2, but the interaction was not significant. However, this result could be confounded by the large number of males in the AVGP group and the large number of females in the non-AVGP group.

Workload

We conducted a 6 (NASA-TLX subscale) × 3 (transparency level) repeated-measures MANOVA on each of the weighted NASA-TLX subscales (Figure 9). The effect of the combined dependent variables (i.e., global workload) was not significant, using Wilks' lambda criterion, $F(12, 18) = 1.14$, $p = .39$, $\eta^2 = .43$, $\lambda = .57$. No differences were found using the univariate ANOVAs among the individual subscales.

Eye Tracking

Due to technical difficulties, we collected eye-tracking data from only 25 participants. We utilized eye movement data as a measure of objective workload. Results of a repeated-measures MANOVA failed to show workload differences across the transparency levels for all of the eye movement variables (mean fixation duration, pupil diameter, saccadic amplitude, or saccade duration).

We found an interaction effect of SpaV on fixation duration and a main effect of SpaO on pupil diameter, $F(2, 46) = 6.19$, $p = .004$, $\eta^2 = .20$

TABLE 1: Operator Performance and Response Time Measures by Transparency Level

Transparency	*M* (*SD*)	*SEM*	95% CI
Correct IA use rate			
Level 1	0.76 (0.15)	0.028	[0.70, 0.82]
Level 1+2	0.87 (0.16)	0.023	[0.83, 0.92]
Level 1+2+3	0.89 (0.12)	0.022	[0.84, 0.94]
Correct IA rejection rate			
Level 1	0.55 (0.20)	0.037	[0.47, 0.62]
Level 1+2	0.67 (0.18)	0.032	[0.60, 0.74]
Level 1+2+3	0.81 (0.20)	0.036	[0.73, 0.88]
Response time (ms)			
Level 1	33.00 (16.73)	3.06	[26.76, 39.25]
Level 1+2	31.53 (18.63)	3.40	[24.58, 38.50]
Level 1+2+3	32.82 (18.51)	3.38	[25.91, 39.73]

Note. CI = confidence interval.

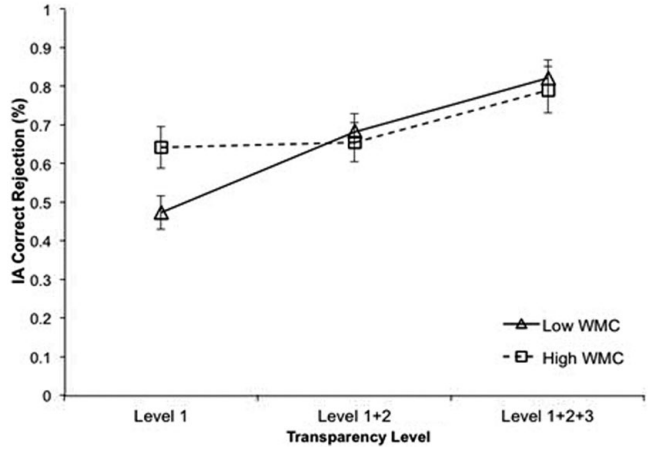

Figure 7. Individual differences between low– and high–working memory capacity (WMC) groups for percentage of intelligent agent (IA) correct rejections across transparency level. Error bars represent standard error of the mean.

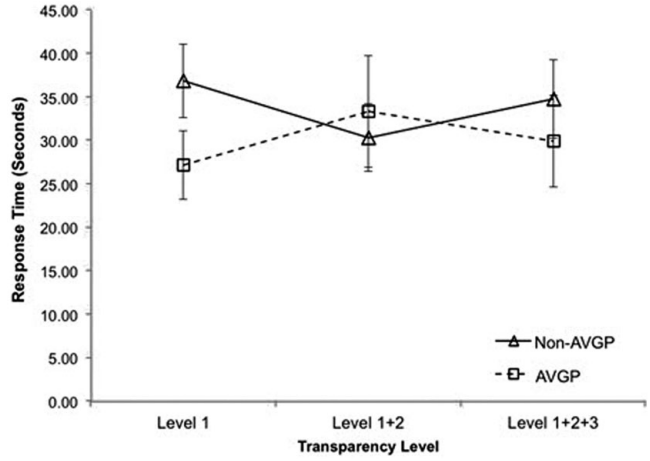

Figure 8. Individual differences in gaming experience among action game players (AVGPs) and non-AVGPs for response time across transparency level. Error bars represent standard error of the mean.

(Figure 10). Individuals with low SpaV had longer fixation durations than high-SpaV individuals in Level 1 (d = 0.42) and Level 1+2 (d = 0.57), but the opposite was observed in Level 1+2+3 (d = 0.51).

Pupil diameter was larger for the high-SpaO group (M = 3.89, SD = 0.17) than for the low-SpaO group (M = 3.50, SD = 0.03) across all transparency levels. The differences were larger in Level 1 (d = 1.04) than in either Level 1+2+3 (d = 0.96) or Level 1+2 (d = .82), $F(1, 23)$ = 5.54, p = .027, η^2 = .19 (Figure 11).

Trust

We conducted two separate between-subjects ANOVAs on the Information Analysis and Decision and Action Selection subscales. Only the trust assessments for the first block were used, as

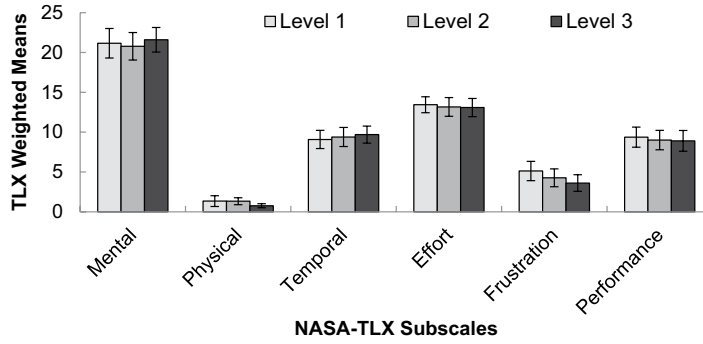

Figure 9. Average weighted NASA Task Load Index subscale means across each transparency level. Error bars are standard error of the mean.

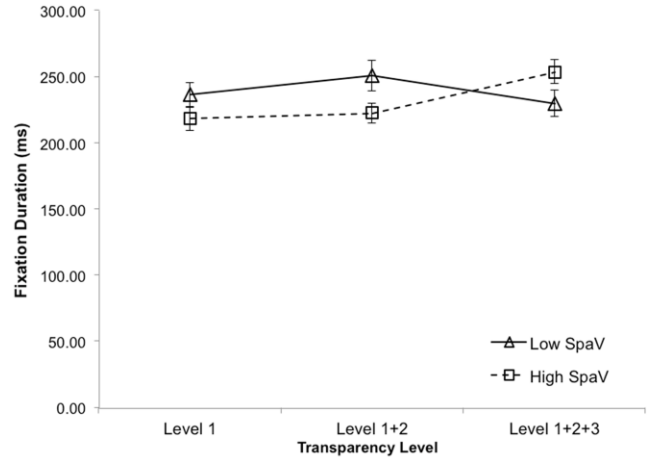

Figure 10. Individual differences in fixation duration between low– and high–spatial visualization (SpaV) groups across transparency level. Error bars represent standard error of the mean.

trust ratings can be biased based on prior experience with the agent (i.e., previous blocks). There were no significant differences across transparency levels for the Information Analysis subscale, $F(2, 27) = 2.14$, $p = .14$, $\eta^2 = .14$ (Table 2).

Results for the Decision and Action Selection subscale were significant for transparency level, $F(2, 27) = 4.01$, $p = .03$, $\eta^2 = .23$ (Table 2). Trust in the system's ability to suggest or make decisions increased as transparency level increased. Post hoc analysis revealed that trust was significantly greater in Level 1+2+3 than in Level 1 ($p = .031$). However, there were no significant differences between Level 1+2+3 and Level 1+2 or between Level 1+2 and Level 1.

System Usability

We conducted a repeated-measures ANOVA on system usability. The analysis revealed a significant effect for transparency level, $F(2, 48) = 5.70$, $p = .006$, $\eta^2 = .11$. Post hoc comparisons indicated that participants found the system more usable in both transparency Level 1+2+3 ($p = .02$) and Level 1+2 ($p = .07$) as compared with Level 1 (Table 2).

DISCUSSION

We investigated the effects of level of agent transparency on operator's task performance, trust, and workload in the context of human–agent

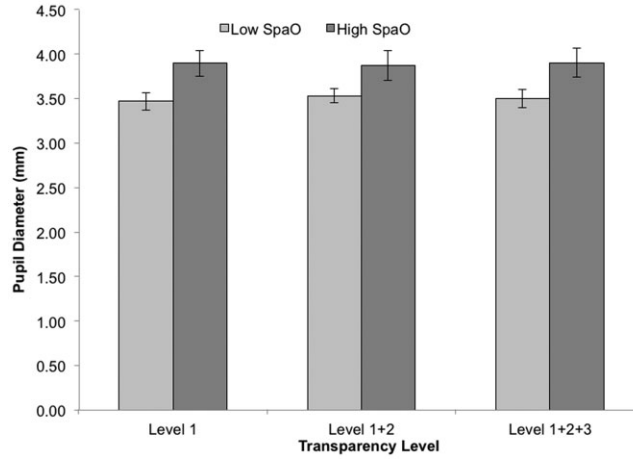

Figure 11. Individual differences in fixation duration between low– and high–spatial orientation (SpaO) groups across transparency level. Error bars represent standard error of the mean.

TABLE 2: Results for Trust Subscales and System Usability Scale (SUS) by Transparency Level

Transparency	*M* (*SD*)	*SEM*	95% CI
Information Analysis trust subscale			
Level 1	5.19 (0.70)	0.22	[4.69, 5.69]
Level 1+2	5.51 (0.73)	0.23	[5.00, 6.03]
Level 1+2+3	5.83 (0.63)	0.20	[5.37, 6.28]
Decision and Action Selection trust subscale			
Level 1	4.63 (0.88)	0.28	[4.00, 5.25]
Level 1+2	4.88 (0.50)	0.16	[4.52, 5.24]
Level 1+2+3	5.47 (0.61)	0.19	[5.03, 5.91]
SUS total			
Level 1	61.83 (20.77)	3.79	[54.08, 69.59]
Level 1+2	66.42 (18.61)	3.40	[59.47, 73.37]
Level 1+2+3	66.75 (19.40)	3.54	[59.51, 74.00]

Note. CI = confidence interval.

teaming for multirobot management. The performance data indicated that participants' correct rejection accuracy increased in relation to transparency level, whereas correct IA usage increased only from Level 1 to Level 1+2. The addition of reasoning information in Level 1+2 increased correct IA use by 11% and correct rejection rate by 12%. The addition of uncertainty information (Level 1+2+3 compared with Level 1+2) improved correct IA use rate by a small amount (2%) and correct rejection rate by 14%.

These results, when combined with response time and workload data, suggest that additional transparency elements in the display improved performance without increasing workload or response time. One alternative explanation for the improved performance with higher level of transparency is that the additional information could have caused participants to be more careful about their decision-making process instead of using the rationale or uncertainty displays. If this explanation is correct, we would expect

response times to increase with higher levels of transparency, but this result was not observed (Table 1). Additionally, participants reported that the sprocket graphic and text tables were the most useful display elements for their decision making. These data (both performance and self-report) suggest that incorporating reasoning and uncertainty information into heterogeneous tactical decision making successfully allowed our participants to make better-calibrated decisions. This finding is consistent with Helldin (2014), although the performance improvement in our study did not come at the cost of longer response times, as Helldin reported.

Previous research indicated that too little or too much trust can lead to disuse or misuse of automated systems (Parasuraman & Riley, 1997). In contrast, our participants who accepted the IA's correct recommendations but (correctly) rejected incorrect recommendations displayed proper trust calibration for the IA's recommendations. Taken together, the correct IA usage and correct IA rejection rates indicate that participants' trust calibration increased as a function of transparency level. This result is consistent with the findings of Oduor and Wiebe (2008), in which transparency improved operator trust calibration and performance in a human–agent joint decision-making context.

Operator performance varied with regard to individual differences in both WMC and gaming experience. Individuals with higher WMC outperformed (correct rejections) those with lower WMC during Level 1. Level 1 provided individuals with only basic information, which in turn required individuals to process and synthesize that information before making a decision. Presumably, individuals with higher WMC more accurately identified incorrect recommendations due to their ability to maintain the information needed to make a judgment in their working memory. Our findings support previous research, which indicates that individuals with low WMC are more likely to rely on heuristics to mitigate their memory load (Quayle & Ball, 2000), but the increase in transparency of the agent interface reduced the performance gap between low- and high-WMC individuals.

Using operator performance as an objective measure of trust reveals only one aspect of the participants' trust in the IA. Participants may have distrusted the system and manually solved each mission without regard to the IA's recommendation (Parasuraman & Riley, 1997). Therefore, we used subjective trust measures to investigate participants' perception of the IA's trustworthiness. Results for the Suggesting or Making Decisions subscale provide evidence that participants trusted the IA's recommendation more when the system was more transparent. Overall, as participants were provided with more transparency, they (correctly) rejected the agent's incorrect recommendation more often yet trusted it more. This increase in trust may be because a more transparent agent—displaying explanations of reasoning and conveying relevant uncertainty—is perceived as more humanlike, intelligent, and trustworthy than an agent whose reasoning is not evident (de Visser et al., 2012; Jian et al., 2000). Another explanation may be that in providing the operator with reasoning and relevant uncertainty information, the IA offers a better understanding of the system's limitations, thus increasing the operator's trust.

This relationship also sheds light on Wickens and Dixon's (2007) findings that automation with less than 70% reliability may not be beneficial to performance. Our results suggest that when the automation's reasoning is transparent, the operator's trust calibration is enhanced, which can result in improved subjective trust and performance even when the automation's reliability is below 70%. Finally, System Usability Scale findings also show the benefit of transparency, as results indicate that participants' perceived usability of the system increased as a function of transparency level.

Prior research has shown that workload and response time can increase as more information is displayed to an operator due to increased information-processing demands, especially when mental workload and effort are high (Helldin, 2014; Zuk & Carpendale, 2007). In contrast, our NASA-TLX and eye-tracking workload data as well as response time failed to show increases as a function of transparency. However, when examining individual differences, we found that participants with low SpaV had longer fixation durations during Level 1 and Level 1+2, which may be a result when these individuals take a

longer time to process the information on screen before making a decision. This finding suggests that spatial abilities may have differential effects on operator response time when agent transparency (especially uncertainty information) is involved but, more importantly, identifies the need for further research. Additionally, in contrast to prior research (Rayner, 2009), we found that pupil diameter was larger for individuals with high SpaO across all transparency levels. However, as Rayner (2009) mentioned, the relationship between pupil diameter and SpaO may change as an experimental task moves closer to simulating a real-world task. Our findings, in conjunction with those of Rayner, suggest that the relationship between pupil diameter and SpaO may be task specific.

Limitations

Taken together, our findings indicate that the benefits of transparency may not necessarily introduce potential costs in terms of speed and workload. Additionally, a more transparent IA engenders higher level of trust and perceived usability. However, the agent behaviors simulated in the current study were based on realistic capabilities of the IMPACT system currently under development, but one key function of IMPACT was purposely removed from the simulation: operator's manual "tweaking" of the agent's recommended plans. Since the goal of the current study was to investigate the effects of transparency levels on operator decisions, greater control over participants' actions (i.e., selecting from two choices rather than proposing their own solutions) was necessary in order to keep variability in participants' responses manageable. Follow-up studies should test the effects of transparency in a more realistic tasking environment and allow the operator to modify the agent's recommended plans. The current study represents our initial effort to demonstrate the potential utility of agent transparency for human decision-making effectiveness.

CONCLUSIONS

Our findings have important implications for the designing of systems to facilitate decision making between the human operator and the IA. In contrast with Helldin (2014), our results indicate that benefits of transparency based on the SAT model can improve performance effectiveness without additional costs in terms of time and workload (Chen et al., 2014).

Additionally, the benefits of transparency were also seen in the participants' trust calibration, which increased as a function of transparency. This result led to participants' feeling the IA was trustworthy and had greater usability when the agent was more transparent. Authors of future research should investigate how incorporating uncertainty into each level of transparency affects operator performance, trust, and workload.

ACKNOWLEDGMENTS

This research was supported by the Department of Defense Autonomy Research Pilot Initiative, under the Intelligent Multi-UxV Planner With Adaptive Collaborative/Control Technologies (IMPACT) project. We wish to thank Isacc Yi, Erica Valiente, Shan Lakhmani, and Jonathan Harris for their contribution to this project. We would also like to thank Gloria Calhoun and Mark Draper for their input.

KEY POINTS

- Increasing agent transparency enhanced the overall human–agent team performance in a multi-unmanned-vehicle management context.
- Operators calibrated their trust in the agent more effectively when the agent was more transparent—their reliance on and compliance with the agent's recommendation was more appropriate.
- The benefits of agent transparency do not necessarily involve either a speed–accuracy trade-off or increased workload costs.
- Operators perceived a transparent agent as being more trustworthy and having greater usability.

REFERENCES

Ahmed, N., de Visser, E., Shaw, T., Mohamed-Ameen, A., Campbell, M., & Parasuraman, R. (2014). Statistical modeling of networked human–automation performance using working memory capacity. *Ergonomics*, *57*, 295–318.

Bevan, N., & Macleod, M. (1994). Usability measurement in context. *Behaviour & Information Technology*, *13*, 132–145.

Bradshaw, J. M., Feltovich, P. J., Johnson, M., Bunch, L., Breedy, M. R., Eskridge, T. C., & Uszok, A. (2008). Coordination in human-agent-robot teamwork. In *International Symposium on Collaborative Technology and Systems* (pp. 467–476). Piscataway, NJ: IEEE.

Brooke, J. (1996). SUS: A quick and dirty usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester, & I. L. McClelland (Eds.), *Usability evaluation in industry* (pp. 189–194). London, UK: Taylor & Francis.

Chen, J. Y., Barnes, M. J., & Harper-Sciarini, M. (2011). Supervisory control of multiple robots: Human-performance issues and user-interface design. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, *41*, 435–454.

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

Chen, J. Y. C., Procci, K., Boyce, M., Wright, J., Garcia, A., & Barnes, M. (2014). *Situation awareness–based agent transparency* (No. ARL-TR-6905). Aberdeen Proving Ground, MD: U.S. Army Research Laboratory.

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, Z. D., Willhelm, O., & Engle, R. (2005). Working Memory Span Tasks: A Methodological Review and User's Guide. *Psychonomic Bulletin & Review*, *12*, 769–786.

Cook, M., & Smallman, H. (2008). Human factors of the confirmation bias in intelligence analysis: Decision support from graphical evidence landscapes. *Human Factors*, *50*, 745–754.

Cummings, M. L., & Mitchell, P. J. (2008). Predicting controller capacity in supervisory control of multiple UAVs. *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans*, *38*, 451–460.

de Visser, E., Krueger, F., McKnight, P., Scheid, S., Smith, M., Chalk, S., & Parasuraman, R. (2012). The world is not enough: Trust in cognitive agents. In *Proceedings of the Human Factors and Ergonomics Society 56th Annual Meeting* (pp. 263–267). Santa Monica, CA: Human Factors and Ergonomics Society.

Douglass, S. (2013). Learner models in the large-scale cognitive modeling initiative. In R. Sottilare, A. Graesser, X. Hu, & H. Holden (Eds.), *Design recommendations for adaptive intelligent tutoring systems learner modeling* (Vol. 1, pp. 111–126). Orlando, FL: U.S. Army Research Laboratory.

Draper, M, H. (2013). *Realizing autonomy via intelligent adaptive hybrid control: Adaptable autonomy for achieving UxV RSTA team decision superiority* [Funding proposal]. Arlington, VA: Office of the Assistant Secretary of Defense for Research and Engineering.

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

Ekstrom, R. B., French, J. W., & Harman, H. H. (1976). *Manual for kit of factor-referenced cognitive tests*. Princeton, NJ: Educational Testing Service.

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors*, *37*, 32–64.

Freedy, E., DeVisser, E., Weltman, G., & Coeyman, N. (2007, May). Measurement of trust in human–robot collaboration. In *Proceedings of IEEE International Symposium on Collaborative Technologies and Systems* (pp. 106–114). Piscataway, NJ: IEEE.

Groom, V., & Nass, C. (2007). Can robots be teammates? Benchmarks in human–robot teams. *Interaction Studies*, *8*, 483–500.

Gugerty, L., & Brooks, J. (2004). Reference-frame misalignment and cardinal direction judgments: Group differences and strategies. *Journal of Experimental Psychology: Applied*, *10*, 75–88.

Hardin, B., & Goodrich, M. A. (2009, March). On using mixed-initiative control: A perspective for managing large-scale robotic teams. In *Proceedings of the 4th ACM/IEEE International Conference on Human–Robot Interaction* (pp. 165–172). New York, NY: ACM.

Hart, S., & Staveland, L. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. Hancock & N. Meshkati (Eds.), *Human mental workload* (pp. 139–183). Amsterdam, Netherlands: Elsevier.

Hegarty, M., & Waller, D. (2004). A dissociation between mental rotation and perspective-taking spatial abilities. *Intelligence*, *32*, 175–191.

Helldin, T. (2014). *Transparency for future semi-automated systems* (Doctoral dissertation). Örebro University, Örebro, Sweden.

Hoffman, J. E., & Subramaniam, B. (1995). The role of visual attention in saccadic eye movements. *Perception & Psychophysics*, *57*, 787–795.

Hwang, S. L., Yau, Y. J., Lin, Y. T., Chen, J. H., Huang, T. H., Yenn, T. C., & Hsu, C. C. (2008). Predicting work performance in nuclear power plants. *Safety Science*, *46*, 1115–1124.

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

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

Levine, T. R., & Hullett, C. R. (2002). Eta squared, partial eta squared, and misreporting of effect size in communication research. *Human Communication Research*, *28*, 612–625.

Lewis, M. (2013). Human interaction with multiple remote robots. *Reviews of Human Factors and Ergonomics*, *9*, 131–174.

Linegang, M., Stoner, H. A., Patterson, M. J., Seppelt, B. D., Hoffman, J. D., Crittendon, Z. B., & Lee, J. D. (2006). Human–automation collaboration in dynamic mission planning: A challenge requiring an ecological approach. In *Proceedings of the Human Factors and Ergonomics Society 50th Annual Meeting* (pp. 2482–2486). Santa Monica, CA: Human Factors and Ergonomics Society.

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). Berlin, Germany: Springer.

Miller, C. A., & Parasuraman, R. (2007). Designing for flexible interaction between humans and automation: Delegation interfaces for supervisory control. *Human Factors*, *49*, 57–75.

Neyedli, H., Hollands, J., & Jamieson, G. (2011). Beyond identity: Incorporating system reliability information into an automated combat identification system. *Human Factors*, *53*, 338–355.

Oduor, K. F., & Wiebe, E. N. (2008, September). The effects of automated decision algorithm modality and transparency on reported trust and task performance. In *Proceedings of the Human Factors and Ergonomics Society 52nd Annual Meeting* (pp. 302–306). Santa Monica, CA: Human Factors and Ergonomics Society.

Paas, F. G., & Van Merriënboer, J. J. (1994). Instructional control of cognitive load in the training of complex cognitive tasks. *Educational Psychology Review*, *6*, 351–371.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*, 230–253.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics–Part A: Systems and Humans*, *30*, 286–297.

Poole, A., & Ball, L. J. (2006). Eye tracking in HCI and usability research. *Encyclopedia of Human–Computer Interaction*, *1*, 211–219.

Quayle, J. D., & Ball, L. J. (2000). Working memory, metacognitive uncertainty, and belief bias in syllogistic reasoning.

Quarterly Journal of Experimental Psychology: Section A, *53*, 1202–1223.

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology*, *62*, 1457–1506.

Riek, L. D. (2012). Wizard of Oz studies in HRI: A systematic review and new reporting guidelines. *Journal of Human–Robot Interaction*, *1*(1).

Russell, S. J., & Norvig, P. (2009). *Artificial intelligence: A modern approach* (3rd ed.) Upper Saddle River, NJ: Prentice Hall.

Sarter, N. B., & Woods, D. D. (1995). How in the world did we ever get into that mode? Mode error and awareness in supervisory control. *Human Factors*, *37*, 5–19.

Spriggs, S., Boyer, J., & Bearden, G. (2014). *Fusion system overview*. Lecture given at Wright Patterson Air Force Base, OH.

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

Wickens, C. D., & Dixon, S. R. (2007). The benefits of imperfect diagnostic automation: A synthesis of the literature. *Theoretical Issues in Ergonomics Science*, *8*, 201–212.

Zuk, T., & Carpendale, S. (2007). Visualization of uncertainty and reasoning. In A. Butz, B. Fisher, A. Krüger, P. Olivier, & S. Owada (Eds.), *Smart graphics* (pp. 164–177). Berlin, Germany: Springer.

Joseph E. Mercado, MSC, USN, PhD, is an aerospace experimental psychologist for the U.S. Navy. He works at the Naval Air Warfare Center Training Systems Division in Orlando, Florida. Prior to his time in the Navy, he worked as a postdoctoral research fellow for the U.S. Army Research Laboratory–Human Research and Engineering Directorate, under Dr. Chen, at the Field Element in Orlando, Florida.

Michael A. Rupp received his MS degree in modeling and simulation in 2012 at the University of Central Florida, where he is pursuing his doctoral degree in the applied experimental and human factors psychology program.

Jessie Y. C. Chen is a senior research psychologist with U.S. Army Research Laboratory–Human Research and Engineering Directorate at the Field Element in Orlando, Florida. She earned her PhD in applied experimental and human factors psychology from the University of Central Florida.

Michael J. Barnes is a research psychologist with the Human Research and Engineering Directorate of the Army Research Laboratory (ARL). Most recently, he was the lead for the human–robotic interaction component of Safe Operations for Unmanned Systems for Reconnaissance in Complex Environments (SOURCE). Prior to ARL, he conducted human factors research (for the U.S. Navy) and served as human factors research manager for the General Electric Corporation. Since coming to ARL, he has served on a number of NATO working groups related to autonomous systems. He has coauthored over 100 articles and coedited a book, *Human–Robot Interactions in Future Military Operations* (2010). His research interests include investigations of risk visualization, intelligence processes, and unmanned aerial vehicles crew systems. He was educated at the University of Arizona (BA) and the New Mexico State University (MA).

Daniel Barber is an assistant research professor at the University of Central Florida's Institute for Simulation and Training. He has extensive experience in the field of robotics and has also developed tools for synchronization, processing, and streaming of data from multiple physiological sensors (e.g., eye tracking, electrocardiogram, and electroencephalography) within simulation and training environments supporting real-time streaming adaptation to user state. His current research focus is on human–robot interaction, including multimodal communication, intuitive user interaction devices, supervisory control of multiple vehicles, and adaptive systems using physiological sensors.

Katelyn Procci, PhD, is a senior human performance engineer with Cubic Corporation. She earned her doctoral degree in human factors at the University of Central Florida, where she worked for the Institute of Simulation and Training in Orlando. During that time, she worked with Dr. Chen's research group and assisted with the early stages of the IMPACT project.

Date received: April 14, 2015
Date accepted: October 26, 2015