The Influence of Agent Reliability on Trust in Human-Agent Collaboration

Xiaocong Fan
School of Engineering
The Behrend College
Pennsylvania State University
Erie, PA 16563, USA
xfan@psu.edu

Sooyoung Oh, Michael McNeese, John Yen
College of Info. Sciences and Tech.
Pennsylvania State University
University Park, PA 16802
soh@ist.psu.edu
mmcneese@ist.psu.edu
jyen@ist.psu.edu

Haydee Cuevas, Laura Strater, Mica R. Endsley
SA Technologies, Inc.
4731 E. Forest Peak
Marietta, GA 30066, USA
haydee.cuevas@satechnologies.com
laura@satechnologies.com
mica@satechnologies.com

ABSTRACT

Motivation – To investigate ways to support human-automation teams with real-world, imperfect automation where many system failures are the result of systematic failure.

Research approach – An experimental approach was used to investigate how variance in agent reliability may influence human's trust and subsequent reliance on agent's decision aids. Sixty command and control (C2) teams, each consisting of a human operator and two cognitive agents, were asked to detect and respond to battlefield threats in six ten-minute scenarios. At the end of each scenario, participants completed the SAGAT queries, followed by the NASA TLX queries.

Findings/Design – Results revealed that teams with experienced human operators accepted significantly less inappropriate recommendations from agents than teams with inexperienced operators. More importantly, the knowledge of agent's reliability and the ratio of unreliable tasks have significant effects on human's trust, as manifested in both team performance and human operators' rectification of inappropriate recommendations from agents.

Originality/Value – It represents an important step toward uncovering the nature of human trust in human-agent collaboration.

Take away message – This research has shown that given even minimal basis for understanding when the operator should and should not trust the agent recommendations allows operators to make better AUDs, to have better situation awareness on the critical issues associated with automation error, and to establish better trust in intelligent agents.

ECCE'08, September 16-19, 2008, Madeira, Portugal.
ACM 978-1-60558-399-0/08/09

Keywords
Reliability, Automation usage decisions, Trust, Agent

INTRODUCTION

Humans and agents are generally thought to be complementary: while humans are more flexible, adaptable, and creative in responding to unforeseen situations, agents are superior in resource-consuming and computation-intensive activities such as information processing, learning, and planning. This has inspired the research on human-centered teamwork (Bradshaw *et al.* 2002; Lennox *et al.* 1999). As a subarea of Multi-agent teamwork (Cohen & Levesque 1991), human-centered teamwork argues for stronger interaction between software agents and their human peers. Within teamwork, both humans and agents are jointly responsible for establishing mutual situation awareness (Endsley 1995), developing shared mental models as situations evolve (Fan & Yen 2007), and adapting to mixed-initiative activities.

While human-centered teamwork promises better overall system performance (e.g., making better decisions that take advantage of information with greater accuracy and finer granularity (Lennox *et al.* 1999; Fan *et al.* 2006)), it could be the other case if inappropriate task allocation policies were adopted or problematic situations emerged due to lack of trust.

Trust is one of the attitudes presented in the "belief-attitude-intention-behavior" sequence; it affects an agent in forming the intention to rely on another agent. Many studies (Parasuraman & Riley 1997; Dzindolet *et al.* 2003; Lee & See 2004) have demonstrated that trust is a meaningful concept to describe human-human interaction and a useful construct to understand human's reliance on automation. For instance, it is reported (Dzindolet *et al.* 2003) that humans tended to distrust an automated decision aid (agent) when observing the agent make errors (exhibit unreliable behavior), and knowing why the aid might err increased their trust in the agent.

From the multi-agent systems perspective, developing trustable technology is a critical factor in the success of agent systems for supporting human-centered teamwork, especially in domains characterized by uncertainty, time-stress, safety and security (e.g.,

command and controls in battlefield). However, despite the abundant research on trust in organizational and sociological disciplines, little practical guidance is available to assist designers in developing trust-aware and trust-adjustable agent systems. Part of the reason is that it is extremely challenging to establish the trust symmetry between humans and agents as it appears in interpersonal trust (in which the trustor and trustee are each aware of the other's behavior and intents (Deutsch 1960)).

In particular, the relationship between trust and agent (automation) reliability merits further investigation in order to address the issue of trust symmetry between humans and agents. Much of the existing research has examined situations in which automation reliability varied, but in ways that seemed random to the human operator. Actual automation reliability varies, but often for well-defined systematic reasons that are comprehensible to human operators. For instance, sensor reliability may vary with atmospheric or weather conditions or with detection distance, which is logical to human operators monitoring such systems. On the other hand, cognitive agents, empowered with naturalistic decision making models (Klein 1997), have been developed and used to support humans in making decisions under time stress (Norling, Sonenberg, & Ronnquist 2000; Norling 2004; Fan *et al.* 2006). However, little research has been aimed to explore the trust issue when cognitive agents act as human's teammates and decision aids. Therefore, in this research, we investigate the influence of systematic failures of a cognitive agent on operators' trust in the agent.

The rest is organized as follows. In Section 2 we review research on trust in automation and introduce the R-CAST agent architecture. Detailed in Section 3 is a task, which is to be used in a human-in-the-loop experiment as described in Section 4 to study the influence of agent reliability on human trust. Section 5 reports experiment results and Section 6 concludes the paper.

BACKGROUND

Trust in Automation

Trust can be defined as the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability (Lee & See 2004). When a person's cognitive resources are not available to support a calculated rational choice due to uncertainty or complexity, trust combining with other attitudes (e.g., effort to engage, self-confidence) will allow the person to form the intention to rely on an agent. For instance, when a human's self confidence is high and trust in an agent is low, he/she is more inclined to override tasks done by the agent.

Trust is a generalized expectancy that is independent of specific interaction experiences and based on the generalization of a large number of diverse experiences (Lee & See 2004). Studies indicate that trust is a dynamic attitude that evolves along with the interaction (Dzindolet *et al.* 2003). For example, a person may initially consider the decision recommendation from an agent trustworthy and reliable. After failure events, his/her trust would decline rapidly and slowly increase again as the agent performs without making errors.

Automation usage decisions (AUDs) can be one of the most important decisions a human operator can make, particularly in time-critical situations. However, biases in trust can lead to misuse or disuse of automation (Parasuraman & Riley 1997; Kaber & Endsley 2004). Misuse refers to over-reliance on automation while disuse refers to neglect or underutilization of automation. For example, in misuse, a war fighter blindly follows the judgments made by an agent; while with disuse, the war fighter either ignores the agent's recommendations or delays action, increasing vulnerability and decision time.

The R-CAST Agent Architecture

The R-CAST agent architecture (Fan & Yen 2007) is built on top of the concept of shared mental models (Cannon-Bowers, Salas, & Converse 1990), the theory of proactive information delivery (Fan, Yen, & Volz 2005), and Klein's Recognition-Primed Decision (RPD) Model (Klein 1997). The R-CAST agent architecture has implemented a "collaborative-RPD" decision process, which is intended to support close human-agent collaborations in relevant information sharing, decision progress monitoring, and expectancy-based decision adaptation.

The RPD model claims that in complex situations human experts usually make decisions based on the recognition of similarities between the current decision situation and previous decision experiences (Klein 1997). R-CAST uses "recognition anchors" to manage the process of decision refinement within a decision space. In particular, an R-CAST agent first starts with the most abstract experience in the current decision space. As more and more information becomes available, potential patterns for further situation evolution become more predictable, and the agent's recognition of a workable experience becomes more and more fine-grained (reaching the lowest possible level of the hierarchy). During the process, an agent also monitors the expectancies associated with the recognized experience. The recognition is reinforced as new events emerge as expected. It is challenged when some expectancy becomes false as the situation evolves; in such a case, the agent can backtrack along the experience hierarchy to seek an experience that is better recognized.

The use of context is of growing importance in developing computational systems that are more responsive to human needs. Specifically, with a better awareness of the decision context, an agent can proactively share information relevant to the context and offer *trustable* intervention/recommendation to its human user. R-CAST distinguishes decision process context, experience context and inference context; these three types of context representation together enable R-CAST to use and integrate various contexts for identifying information relevant to decision making, for adapting decisions to a dynamic environment, and for facilitating reuse of context-related domain knowledge.

R-CAST offers two approaches to achieve teamwork adaptability. First, it supports a richer structure of a functional SMM that not only covers team structures and team processes used by agents to infer collaboration needs, but also covers dynamic teamwork contexts. Such an enhanced representation of shared mental models enables an R-CAST agent to better manage its own 'focal' attention and to initiate human-agent collaboration in a human-appreciatable way. Second, non-trivial collaborative multi-agent systems need to continuously make decisions, which demands a group of agents to coordinate not only on domain-specific tasks but also in the decision-making process itself. Meshing humans' decision making process with agents' decision making process promises better human-agent collaboration. It, however, requires humans and agents to maintain a shared understanding of the decision making progress. To achieve this, R-CAST has incorporated a naturalistic decision making process (RPD) that may well support human decision makers better than more arbitrary rule-based systems.

R-CAST has been employed as teammates and decision aids (Fan *et al.* 2006) of Command and Control (C2) human operators, helping address the informational challenges in team decision making under stress in a simulated battlefield environment. While the result indicated that R-CAST agents can significantly improve the tasking capacity of C2 teams in time-stressed situations involving multiple decision contexts, the study also left open the question of human-agent trust: What factors might have impacts on a human's trust (and use) of his/her decision aids? We here take a step toward this direction, investigating how human's trust might be affected by *imperfect* cognitive agents.

TASK DESCRIPTIONS

We have implemented a simulation environment called "Three-Block Challenger", where in an urban area a command and control team has to frequently conduct humanitarian, peacemaking and combat missions in close proximity (e.g., within three blocks). It imposes challenging information and decision making demands associated with the command and control of urban operations.

The synthetic task environment can produce three types of threats: Improvised Explosive Device (IEDs), crowds, and insurgents, which represent the targets of humanitarian, peacekeeping, and combat operations, respectively. *IEDs* are motionless targets, and if exploded, can cause damage to the nearby objects. A *crowd* represents a group of people which may contain activists that can be friends or foes. A crowd can be of medium (M) or large (L) size, and the group size of a crowd can change over time. Two crowds can merge together if they move close enough. Another type of movable targets is *insurgents*, each is associated with a threat level that can be L(low), M(medium), or H(high).

Other objects of interest in the environment are main supply routes (MSRs) and key buildings (religious buildings, schools, and hospitals). There are also limited number of friendly units, squads and Explosive Ordnance Disposal (EOD) teams, under the control of a C2 team.

In this study, a C2 team consists of an S2 suite (intelligence cell) and an S3 suite (operations cell). The roles of C2 operators have been simplified. S2 is responsible for processing incoming reports, called Spot reports; collecting relevant information from other sources; and alerting S3 of potential threats. S3 needs to process alerts from S2, and make decisions on which target to handle next and which resources (friendly units) to allocate toward that target.

Table 1: Requirements on handling targets

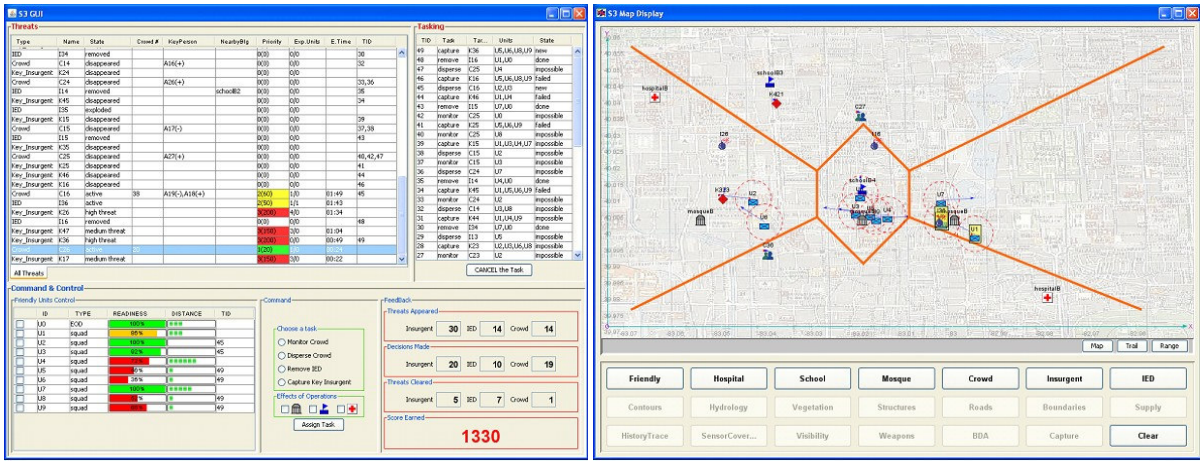
Targets			*Value*	*Res. Req.*	*Action*
Crowd	M	w/o foe	20	1U	monitor
	M	w foe	40(+10)*	2U	disperse
	L	w/o foe	40(+10)*	2U	disperse
	L	w foe	50(+10)*	3U	disperse
Insurgent (3 threat levels: L, M, H)			n=1,2,3 for L,M,H		
			50+50n	(n+1)U	capture
IED			60(+20)*	1U + 1E	remove

'U' refers to "squad unit", 'E' refers to EOD team.
**additional credit value when a target is near an MSR.*

Decision making in target selection and resource allocation requires the S3 suite to consider trade-offs among multiple factors: target type, threat level, the combat readiness of the available units, the unit-target distance, and staying time of each active target (how long it has been on the field). The type and threat level of a target determine how many friendly units will be needed to handle the target. Table 1 lists for each type of target the credit value (the reward points a C2 team can get if a target is handled successfully), the number of resources required to handle a target, and what action S3 should take. For example, the second entry says that dispersion of a medium-sized crowd with a foe needs two squad units, and 40 points can be credited if the crowd is dispersed successfully. The last entry says that one squad unit and one EOD team are required to remove an IED. If successful, 60 points can be credited if the IED is close to buildings only or MSRs only, 80 points if it is close to both.

The combat readiness of a friendly unit, represented by a percentage value, indicates how well the unit has prepared for handling threats. The readiness value decreases by a certain amount after a unit is applied to a threat, and can recover incrementally as time passes.

A target may appear, stay on, and disappear from the battle field following certain temporal and spatial patterns unknown to human operators. The staying time of a target and the distance from the available units to the target affect decision making in target selection: Assuming a Poisson model of lifespan, a target that has a longer staying time or is farther away from the available units should not be selected first due to less chance of mission success.



(a) (b)

Figure 2: (a) The human-agent interaction display of S3 suite; (b) The map display of S3 suite.

METHODOLOGY

The specific objective of this study is to determine whether human operators' *a priori* experience and their knowledge of agent reliability affect their trusts on a cognitive agent as reflected in their appropriate use of the decision recommendations from the agent.

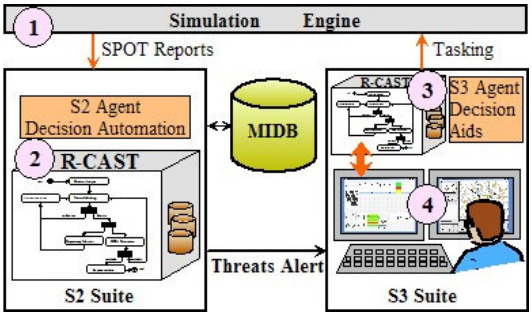


Figure 1: The environment

Environment

As shown in Figure 1, the experiment environment involves a Simulation Engine, S2 suite, and S3 suite, with a human operator in the loop.

At each cycle, the Simulation Engine produces Spot reports for all the active targets and friendly units on the field and sends the reports to the S2 suite. The role of S2 suite is played by an R-CAST agent (S2 agent) equipped with decision making experience drawn from domain experts. S2 agent interacts with the simulated MIDB (Military Intelligence Database) to gather relevant information, recognizes potential threatening targets on the field and alerts the S3 suite of the threats.

The role of S3 suite is played by an R-CAST agent (S3 agent) and a human operator. The human operator has equipment with two monitors: a map display for tracking situation development, and a graphical user interface (GUI) for collaborating with S3 agent to handle threats.

The interaction display (Fig. 2(a)) consists of a threats table, a tasking table, a unit control table, a command panel and a feedback display panel. The threats table shows consolidated information of the threats on the field: for each threat, it gives threat type, ID, status, crowd size, activists associated with a crowd, nearby buildings, priority, requirements on friendly units, elapsed time, and the IDs of the tasks against the threat if applicable. Threat priorities are color-coded: green, yellow, and red represent low, medium, and high threatening targets, respectively. Once removed, higher threatening targets also contribute higher rewarding points to the performance.

The tasking table shows task details, including the target under concern, the assigned friendly units, and the task status, which can be 'new', 'done', 'impossible', or 'failed'. A task is futile (i.e., no reward points) if it becomes impossible; this happens when a target disappears before it is surrounded by the assigned friendly units. A task fails if the assigned friendly units have insufficient combat readiness when they fire at the target. The unit control table lists all the resources, the corresponding combat readiness, and their respective distance to the selected target. The units engaged in a task are available for reuse once the task is done, failed or becomes impossible.

After resource allocation, the command panel allows the S3 human operator to 'physically' issue a task applicable to the target being selected. To engage participants in a more realistic decision-making situation, a secondary task is present where the S3 human operator, when sending out friendly units, needs to check the key buildings nearby the selected target. This means that an operator has to maintain attention to both the interaction display and the MapDisplay.

The feedback display panel shows some statistics about tasks issued, threats cleared, and reward points earned.

In the experiment, the S3 agent offers decision aids and recommendations to the S3 operator. It is, however, the S3 operator who has the final authority for decisions on target selection and resource allocation.

The MapDisplay (Fig. 2(b)) shows the snapshot (refreshed every 5 s) of all the active entities on the field. It allows a human operator to highlight the target of interest and trace its movement, to figure out the

spatial relationships among threats and friendly units, and to project forward the location of a moving targets. As we mentioned above, a human operator needs to check the key buildings nearby the threat to be handled. Determining the key buildings nearby a threat simply based on the snapshot can be wrong because the snapshot is updated with incoming Spot reports every 5 seconds whereas the threat might have moved to a different location in the meantime. The moving direction indicator associated with a target in the MapDisplay can be very useful to better determine nearby key buildings.

EXPERIMENTAL CONDITIONS

This experiment involved four factors: Population Group (PG), Knowledge of Agent Reliability (KAR) with two levels 'known' and 'unknown', Task Complexity (TC), and reliability level of agent recommendations (RIT).

Subjects

Thirty university students (23 males and 7 females) majored in Information Science and Technology (IST) with average video-game experience 5.8 hours per week were recruited as one group. Another thirty participants (28 males and 2 females) were recruited from a US Army ROTC (Reserve Officer Training Corps) organization with average video-game experience 3.7 hours per week. Thus, two levels (ROTC vs. IST) of PG were considered.

Systematic Errors of Agent

The collaboration between S3 operator and S3 agent is critical to the overall performance especially when the human operator is cognitively overloaded under high time stress. Whenever S3 operator selects a target, S3 agent helps make everything ready for tasking: check the nearby buildings if applicable, and allocate sufficient number of units that are optimal relative to certain constraints.

As we mentioned before, decision making in resource allocation requires the consideration of trade-offs among multiple factors. To introduce *systematic errors* to the S3 agent so that the reliability of agent recommendations can be manipulated, we purposely configured S3 agent such that it makes recommendations regarding resource allocation without considering the combat readiness of friendly units.

In particular, we designed the experiment scenarios such that only the insurgent tasks need the consideration of combat readiness: If the combat readiness is lower than 80% threshold for any unit assigned to capture a key insurgent, the task will not succeed. Consequently, agent recommendations regarding crowd and IED tasks are always reliable while the agent makes systematic errors for insurgent tasks.

Scenario Design

We designed 6 scenarios to vary the other two independent variables: tasking complexity and reliability level. In this experiment task complexity is characterized by the number of active targets on the field: the situation is more demanding when there are more active targets. We defined two levels of tasking complexity: M (with 8 active targets) and H (with 12 active targets). Since a target can be removed or disappear by itself, to ensure that there are desired number of active targets on the field, the scenarios were designed such that the disappearance of one target will trigger a new target to pop up.

The reliability level was controlled by varying the ratio of the insurgent tasks (RIT) to the total number of tasks. Three ratios were considered: 1/4, 1/3, and 1/2. For instance, a scenario with ratio 1/3 means among all the targets ever appeared on the field, 1/3 of them are insurgents. Thus, in total we designed 6 scenarios reflecting the different combinations of TC and RIT levels. The scenarios were randomized in the settings of initial locations (targets, MSRs, key buildings, IEDs), targets' appearance time, waypoints and velocities of movable targets, sizing of crowds, and threat levels of insurgents. Each scenario lasted 10 minutes.

In sum, this is a mixed $2 \times 2 \times 2 \times 3$ factorial treatment design (PG × KAR × TC × RIT), where TC (task complexity) and RIT (ratio of insurgent threats) are within-subjects variables and KAR (knowledge of agent reliability) is a between-subjects variable. Fifteen replications of each response were collected for each treatment condition.

Procedures

Participants were initially familiarized with the domain tasks and the environment settings, then went through a training session. The instructions to participants included detailed descriptions of both the MapDisplay and the interaction display, as well as where the S3 agent can offer decision recommendations and how to accept/adjust agent recommendations. His/her goal is to maximize the reward points by removing as many threats as possible.

The training session was also used to group participants (randomly) into two levels of "knowledge of agent reliability." For participants belonging to the 'known' group, they were told that "agent recommendations are highly reliable; the sole source of agent unreliability is combat readiness--the agent does not consider combat readiness in its recommendations." For participants belonging to the 'unknown' group, they were told that "the reliability of agent recommendations is unknown." However, regardless of condition, all participants were informed of the rule "If combat readiness is less than 80% for any unit assigned to capture an insurgent, the task will not succeed."

After the training session, the participants were permitted to practice the tasks for 10 min. After a 5-min break, all participants were required to complete six 10-min trials, each followed by the completion of the SAGAT queries, followed by the NASA TLX queries. For each participant, the six scenarios were scheduled in random orders.

Dependent Measures

Several response variables were measured on the 360 trials of the experiment (180 runs for the 30 ROTC participants (expert operators) and 180 runs for the 30 IST participants (novice operators)).

For each experiment run i, we recorded n_{aKi}, n_{bKi}, and n_{cKi} —the numbers of key-insurgents captured with high, medium, and low threats respectively; n_{aDi}, n_{bDi}, and n_{cDi}—the numbers of IEDs removed with high, medium, and no threats respectively; and n_{aCi}, n_{bCi}, n_{cCi}, and n_{dCi}— the numbers of crowds dispersed with high, slightly high, medium, and low threats respectively. Let $n_{Ki} = n_{aKi} + n_{bKi} + n_{cKi}$, $n_{Di} = n_{aDi} + n_{bDi} + n_{cDi}$,

$n_{Ci} = n_{aCi} + n_{bCi} + n_{cCi} + n_{dCi}$.

The Average Performance Index (API) is defined as:

$API_i = (\Sigma_{X\in\{K,D,C\}}(\theta_{Xi}/n_{Xi}))/(n_{Ki} + n_{Di} + n_{Ci})$; where

$\theta_{Ki} = 200n_{aKi} + 150n_{bKi} + 100n_{cKi}$,

$\theta_{Di} = 80n_{aDi} + 60n_{bDi} + 0n_{cDi}$, and

$\theta_{Ci} = 60n_{aCi} + 50n_{bCi} + 40n_{cCi} + 20n_{dCi}$,

where the weights in computing θ_{Ki}, θ_{Di}, θ_{Ci} are the credit values of the corresponding threatening targets. The API measure reflects a team's overall performance (competency).

Also measured are IRA (Inappropriate Recommendations Accepted) and IRC (Inappropriate Recommendations Correctly adjusted); these two measures are closely related to a human operator's trust and reliance on the decision recommendations from the imperfect agent. Intuitively, the more trust an operator has on an agent, the more likely (number of times) that he/she 'blindly' accepts the inappropriate recommendations from the agent, and the less likely that he/she intends to correct the inappropriate recommendations. For the domain problem as described above, this means that an operator with no knowledge of the agent reliability might mistakenly accept recommendations regarding insurgent tasks (until he/she discovers the systematic errors made by the agent) more times than an operator who knows the agent reliability prior to a trial.

RESULTS

Task Performance

We conducted four-way analysis of variance (ANOVA), with the significance level $\alpha = 0.05$ adopted.

The ANOVA output indicates that the ratio of insurgent threats (RIT) had significant effects on the performance as measured by API. Fig. 3(a) gives the Boxplot of API as RIT varies. As the ratio of insurgent threats increased from 1/4 to 1/3, the C2 performance index improved (the mean increased from 1.51 to 1.76), while as RIT increased to 1/2, the performance dropped significantly (the mean value dropped to 1.387).

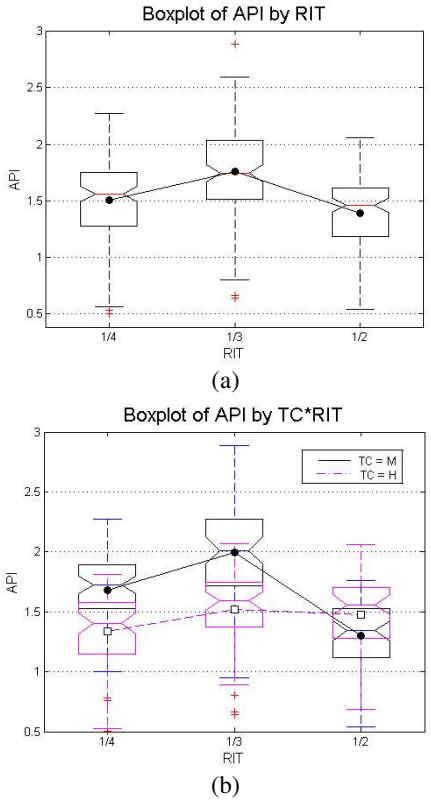
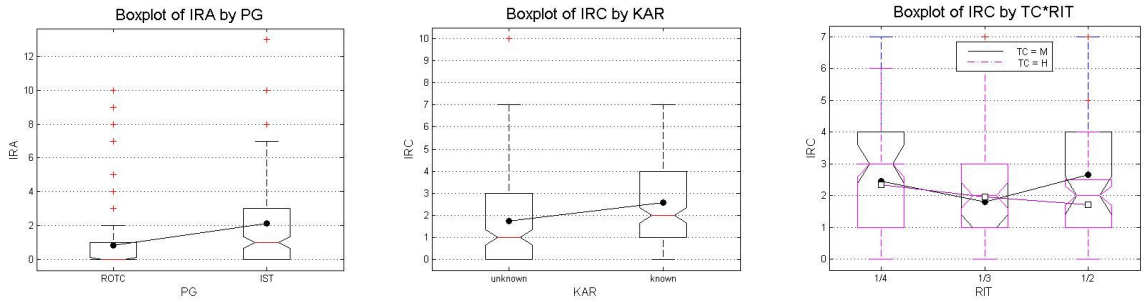


Figure 3: Team performance

Further scrutiny of the data allowed us to find that when RIT was 1/4, the percentage of successful insurgent tasks was 62.27% (out of the 11.992 average number of attempts), while this percentage increased to 65.5% (out of 10.7 average number of attempts) when RIT was 1/3. Moreover, 69.21% of the successful insurgent tasks were high-level insurgent threats under the 1/3 RIT condition, as compared to 57.59% under the 1/4 RIT condition. Thus, the performance increase as RIT changed from 1/4 to 1/3 can be attributed to the fact that the 1/3 RIT condition may present a more favorable situation to the S3 operators, who were able to pay more careful attention to the insurgent threats (attempted less but attacked the keys). However, as more insurgent threats were present when RIT became 1/2, the S3 operators on average attempted on 13.617 insurgent tasks, of which only 57.46% succeeded. The performance dropped because the S3 operators became cognitively overloaded when they confronted with too many insurgent threats. This can be revealed by the irrational attention allocation over the three types of threats. When RIT was 1/3, the attention was paid equally to insurgents, IEDs, and crowds (with 10.7, 10.23, and 9.04 tasks attempted, respectively), while the attention was paid too much on insurgent tasks when



(a) (b) (c)

Figure 4: Agent reliability on human trust

RIT was 1/2 (with 13.62, 9.72, and 7.23 tasks attempted, respectively). The consequence is threefold. When RIT changed from 1/3 to 1/2, (1) the percentage of successful crowd tasks dropped from 38.63% to 27.57% due to less attention paid; (2) the percentage of successful insurgent tasks dropped from 65.5% to 57.46%, while the percentage of futile tasks (attempted but unfinished) increased from 15.89% to 26.93%; and (3) although the percentage of successful IED tasks increased from 58.22% to 66.89%, 5.5% of which were actually futile (attacked harmless IEDs with no reward point), which, again, could be attributed to the unbalanced attention.

It also indicates that there are two-way interactions between RIT levels and TC levels, as shown in Fig. 3(b). While the above analysis still applies here, it is clear that the C2 performance were much better when TC is 'M' than when TC is 'H' in situations with medium (1/3) or low (1/4) RIT. However, the opposite is true when RIT is 1/2: the S3 suite had the worst performance when TC='M' and RIT=1/2.

Further data analysis revealed that when RIT is 1/2, the S3 operators could balance their attention much better under the condition TC='H': the attempted tasks on insurgent, IED, and crowd threats had the distribution (38%, 39%, 23%), as compared to (49.8%, 25.9%, 24.3%) when TC='M'. The consequence is that when TC='M', the S3 operators wasted resources on about 40% futile insurgent tasks and 12.7% futile IED tasks, and only about 21% crowd tasks were successful. One interpretation is that, when there were more insurgents, the S3 operators were allured to attack more insurgent threats, with most of the tasks however, turned out to be futile (the targets disappeared before being surrounded by the assigned units); while when there were too many insurgents, the operators simply changed their strategy: instead of wasting resources on insurgents beyond their capacity, they distributed the limited resources (time, units, and cognition) to all the three types of threats. In sum, it seems that the S3 operators tend to favor the 1/3 RIT condition, and given the limited resources, they may not be able to balance their attention appropriately when the RIT is too high (1/2).

Agent Reliability on Human Trust

The S3 operators' trust and reliance on the S3 agent can be revealed by analyzing the IRA (Inappropriate recommendations accepted) and IRC (Inappropriate recommendations adjusted correctly) responses.

The ANOVA output indicates that both population group (PG) and ratio of insurgent threats (RIT) had significant effects on IRA. As RIT increased, the S3 operators accepted more inappropriate recommendations from the S3 agent. More interestingly, as shown in Fig. 4(a), the ROTC operators accepted far less inappropriate recommendations than the IST operators. In other words, the novice operators tended to rely more on the agent than the operators who had a priori C2 operation experience.

The ANOVA output also indicates that both the knowledge of agent reliability (KAR) and ratio of insurgent threats (RIT) had significant effects on IRC. As RIT increased, the S3 operators correctly adjusted less inappropriate recommendations from the S3 agent. More interestingly, knowing the agent reliability helped the S3 operators rectify more number of inappropriate recommendations (Fig. 4(b)). This seems to suggest that, with the knowledge of agent reliability, participants had more trust on the agent.

This concurs with the NASA TLX trust survey conducted immediately after each participant finished a trial. The survey responses were recorded on a 7-point rating scale, ranging from 1 (agree) to 7 (disagree). The data indicates that the participants without the knowledge of agent reliability tended to agree that the decision aid is deceptive ($\mu = 3.67$; $\delta = 1.47$), while the participants with the knowledge tended to be neutral or disagree ($\mu = 4.57$; $\delta = 1.55$).

There also exist two-way interactions between RIT levels and TC levels on IRC, as shown in Fig. 4(c). As RIT increased, while the S3 operators could rectify less inappropriate recommendations under high task complexity, under medium task complexity there was a big improvement as RIT changed from 1/3 to 1/2. Take the API performance (Fig. 3(b)) into consideration, while on average the condition where RIT=1/2 and TC=M allowed the S3 operators to rectify the most number of inappropriate recommendations, it produced the worst overall performance. One interpretation is that probably the S3 operators had paid too much attention on the insurgent tasks. Although the S3 operators had

adjusted the unit allocation appropriately, many of which, however, turned out to be futile.

CONCLUSION

As the need for human-centered multi-agent systems increases (in domains such as battlefield, healthcare), trust will become increasingly important for mediating human-agent interactions involving uncertainty, security, and reliability.

In this research, we investigated several factors surrounding the challenging problem of human trust on cognitive agents with varying levels of reliability caused by systematic errors. The experiment represents an important step forward in uncovering the nature of human trust in human-agent collaboration. The result demonstrated that while experts tend to be cautious, novice people tend to take more advantage of the agent reliability knowledge, and rely more on the agent recommendations in tasking. More importantly, it suggested that given even minimal basis for understanding when the operator should and should not trust the agent recommendations allows operators to make better AUDs, to have better situation awareness on the critical issues associated with automation error, and to establish better trust in intelligent agents.

This study also reveals that people can easily become cognitively overloaded in high demanding situations, and it is desirable to develop 'trust-aware' agent technologies such that an agent could develop/learn trust models of its human users over time, and offer adjustable autonomy by monitoring users' interaction attitudes and reliance patterns.

REFERENCES

Bradshaw, J.; Sierhuis, M.; Acquisti, A.; Gawdiak, Y.; Prescott, D.; Jeffers, R.; Suri, N.; and van Hoof, R. 2002. What we can learn about human-agent teamwork from practice. In *Workshop on Teamwork and Coalition Formation, Autonomous Agents and Multi-agent Systems (AAMAS 02)*.

Cannon-Bowers, J. A.; Salas, E.; and Converse, S. 1990. Cognitive psychology and team training: Training shared mental models and complex systems. *Human Factors Society Bulletin* 33:1–4.

Cohen, P. R., and Levesque, H. J. 1991. Teamwork. *Nous* 25(4):487–512.

Deutsch, M. 1960. The effect of motivational orientation upon trust and suspicion. *Human Relations* 13:123–139.

Dzindolet, M. T.; Peterson, S. A.; Pomranky, R. A.; Pierce, L. G.; and Beck, H. P. 2003. The role of trust in automation reliance. *International Journal of Human-Computer Studies* 58:697–719.

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

Fan, X., and Yen, J. 2007. R-CAST: Integrating team intelligence for human-centered teamwork. In *Proceedings of the Twenty-Second National Conference on Artificial Intelligence (AAAI'07)*, 1535–1541.

Fan, X.; Sun, B.; Sun, S.; McNeese, M.; and Yen, J. 2006. RPD-enabled agents teaming with humans for multicontext decision making. In *Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems*, 34–41. ACM Press.

Fan, X.; Yen, J.; and Volz, R. A. 2005. A theoretical framework on proactive information exchange in agent teamwork. *Artificial Intelligence* 169:23–97.

Kaber, D. B., and Endsley, M. R. 2004. The effects of level of automation and adaptive automation on human performance, situation awareness and workload in a dynamic control task. *Theoretical Issues in Ergonomics Science* 5(2):113–153.

Klein, G. A. 1997. The recognition-primed decision (rpd) model: Looking back, looking forward. In *Naturalistic decision making (Eds: C. E. Zsambok and G. Klein)*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc. 285–292.

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

Lennox, T. L.; Payne, T.; Hahn, S. K.; Lewis, M.; and Sycara, K. 1999. MokSAF: How should we support teamwork in human-agent teams? Technical Report CMU-RITR- 99-31, Robotics Institute, Carnegie Mellon University, PA.

Norling, E.; Sonenberg, L.; and Ronnquist, R. 2000. Enhancing multi-agent based simulation with human-like decision making strategies. In Moss, S., and Davidsson, P., eds., *Proceedings of the Second International Workshop on Multi-Agent Based Simulation*, 214–228.

Norling, E. 2004. Folk psychology for human modelling: Extending the BDI paradigm. In *AAMAS '04: International Conference on Autonomous Agents and Multi Agent Systems*, 202–209.

Parasuraman, R., and Riley, V. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human Factors* 39(2):230–253.