Trust and Reliance on an Automated Combat Identification System

Lu Wang and **Greg A. Jamieson,** University of Toronto, Toronto, Ontario, Canada, and **Justin G. Hollands,** Defence Research and Development Canada, Toronto, Ontario, Canada

Objective: We examined the effects of aid reliability and reliability disclosure on human trust in and reliance on a combat identification (CID) aid. We tested whether trust acts as a mediating factor between belief in and reliance on a CID aid. **Background:** Individual CID systems have been developed to reduce friendly fire incidents. However, these systems cannot positively identify a target that does not have a working transponder. Therefore, when the feedback is "unknown", the target could be hostile, neutral, or friendly. Soldiers have difficulty relying on this type of imperfect automation appropriately. **Method:** In manual and aided conditions, 24 participants completed a simulated CID task. The reliability of the aid varied within participants, half of whom were told the aid reliability level. We used the difference in response bias values across conditions to measure automation reliance. **Results:** Response bias varied more appropriately with the aid reliability level when it was disclosed than when not. Trust in aid feedback correlated with belief in aid reliability and reliance on aid feedback; however, belief was not correlated with reliance. **Conclusion:** To engender appropriate reliance on CID systems, users should be made aware of system reliability. **Application:** The findings can be applied to the design of information displays for individual CID systems and soldier training.

INTRODUCTION

Individual Combat Identification Systems

Throughout history, the difficulty of distinguishing friend from foe on the battlefield has contributed to friendly-fire incidents (Gimble, Ugone, Meling, Snider, & Lippolis, 2001; Jones, 1998). Technical solutions have the potential to improve soldiers' combat identification (CID) ability. A recent example is the individual CID system, which is intended to help dismounted infantry soldiers identify other friendly soldiers, particularly in urban operations ("Friend or Foe," 2007; Lowe, 2007; K. Sherman, 2000; K. B. Sherman, 2002).

The CID system consists of an interrogator and a transponder (Boyd et al., 2005). A soldier equipped with an interrogator directs an encrypted laser query at an unidentified soldier. If the unidentified soldier is wearing a proper transponder, it will decode and validate the interrogation message and send a reply. If the interrogator receives the correct reply, a light-emitting diode on the weapon blinks to give "friend" feedback. Otherwise, no feedback is given ("Friend or Foe," 2007; K. Sherman, 2000), which can be seen as implicit "unknown" feedback.

The feedback is "unknown" instead of "enemy" because the interrogator cannot positively identify enemies. The target could be a civilian, a neutral soldier, an enemy, or a friendly soldier who does not have a working transponder (Snook, 2002). The reliability of "unknown" feedback from the CID system—the probability that a target is hostile given "unknown" feedback—can vary moment by moment in the battlefield. However, it is an extremely rare event that a CID system reports "friend" without a

Address correspondence to Greg A. Jamieson, Department of Mechanical and Industrial Engineering, University of Toronto, 5 King's College Road, Toronto, Ontario, M5S 3G8, Canada; jamieson@mie.utoronto.ca. ***HUMAN FACTORS,*** Vol. 51, No. 3, June 2009, pp. 281-291. DOI: 10.1177/0018720809338842.

coded reply from a transponder (Karsh, Walrath, Swoboda, & Pillalamarri, 1995). Thus, the likelihood that the system gives "friend" feedback for a hostile target is close to zero.

Automation Reliability, Trust, and Reliance

The benefit of CID systems depends on whether soldiers can rely on the automation feedback appropriately. Successful reliance strategies depend, in part, on operators having appropriate trust in the automation (Lee & Moray, 1992; Lerch, Prietula, & Kulik, 1997; Madhavan & Wiegmann, 2004; Masalonis & Parasuraman, 2003; Muir & Moray, 1996). In general, studies with a variety of automated systems have shown that trust in, and reliance on, imperfect automation can be appropriate when automation reliability information is made available to the user (Cohen, Parasuraman, & Freeman, 1998; Lee & See, 2004; Sheridan & Parasuraman, 2006; St. John, Smallman, Manes, Feher, & Morrison, 2005). For example, St. John et al. (2005) had participants identify and respond to significant threats on a naval tactical display. An imperfect algorithm decluttered the display by dimming the less threatening symbols. Participants informed of the fallibility of the aid appropriately distrusted its assessment and manually checked the dimmed symbols.

In general, human–automation performance improves with automation reliability (Wickens & Dixon, 2007). In contrast, in existing studies on CID systems, even high-reliability CID aids failed to improve target identification accuracy (Dzindolet, Pierce, Beck, Dawe, & Anderson, 2000, 2001; Dzindolet, Pierce, Pomranky, Peterson, & Beck, 2001; Karsh et al., 1995; Kogler, 2003). Participants in these studies did not rely on automated feedback appropriately even when the aid reliability information was disclosed. Because trust in the CID systems was not measured in these studies, it is difficult to attribute the improper reliance to inappropriate trust or other factors (Parasuraman & Mouloua, 1996). In some of the conditions in these studies (Dzindolet et al., 2000; Dzindolet, Pierce, Beck, et al., 2001; Dzindolet, Pierce, Pomranky, et al., 2001; Kogler, 2003), the levels of automation reliability were relatively low (< .70), which might have led to distrust (Wickens & Dixon, 2007). In one study (Karsh et al., 1995), users had to wait for feedback from the automated system while needing to respond quickly; the cost of reliance may have led to disuse of the automation. However, taken in combination, the lack of appropriate reliance on feedback from imperfect CID aids across multiple studies raises the question of whether CID is somehow unique in the automation domain.

The CID automation context does appear to differ in important ways from automation in other contexts. In supervisory control, the automation typically supplants the human in a task that is performed concurrently with a higher-priority (often manual) task. In contrast, CID automation supports the user in performing a single perceptual task. In CID, the user is actively engaged in the target identification task. However, in supervisory control, the human user often does not attend to an event until alerted by an automated system. Furthermore, as noted earlier, CID automation tends to fail one way and not another. An automation miss (i.e., "friend" feedback for a hostile target) is much less likely than an automation false alarm (FA; i.e., "unknown" feedback for a friendly target). Finally, considering the life-or-death consequences of virtually every decision in combat, a user's arousal levels are likely very high.

Given these differences, it is possible that the results found in other automation contexts may not directly apply to the CID situation. However, it is also possible that the measurement of reliance in earlier CID studies did not permit a proper understanding of the reliance. We examine this question in the next section.

Measuring Reliance

Wang, Jamieson, and Hollands (2008) identified two common approaches to measuring reliance:

1. Measure the consistency or correlation between automation feedback and the user's decision, assuming that the greater the consistency, the greater the reliance on automation (e.g., Biros, Daly, & Gunsch, 2004; Bisantz & Pritchett, 2003; Brunswik, 1956; Murrell, 1977). This consistency

approach is related to Brunswik's lens model (1956), which posits that if a judgment task is based on multiple cues, the reliance on each cue can be inferred by the correlation between each cue and the final judgment. The disadvantage of this approach is that the consistency between a user's decision and the automation feedback may not be caused by the user's reliance per se. For example, if both the user and the automation are highly accurate in performing a task, the consistency between them will be high regardless of reliance.

2. Estimate the reliance by comparing users' misuse and disuse rates (e.g., Dzindolet, Pierce, Beck, et al., 2001; Dzindolet, Pierce, Pomranky, et al., 2001; Parasuraman & Riley, 1997). Users who are more likely to misuse than disuse automation can be deemed to have shown reliance. When misuse exceeds disuse, the intuition might be that the user overrelies on the aid. However, without considering the relative accuracy of the automation and the user's unaided performance, this conclusion may not be correct.

Wang et al. (2008) argued that neither of these reliance measures provides a clear definition of *optimal reliance*, without which the judgment of the appropriateness of reliance can be ambiguous. Instead, they favored an approach based on signal detection theory (SDT; Green & Swets, 1966; Macmillan & Creelman, 1991; Wickens & Hollands, 2000). Two indicators—sensitivity and response bias—characterize an observer's responses in a signal detection task. Both measures are derived from two scores: hit rate, P(Hit), and FA rate, P(FA). Importantly, an observer's optimal response bias can be defined in any given condition with a predetermined signal rate and decision payoff structure. Researchers have used response bias to indicate reliance on a binary warning indicator (e.g., Maltz & Meyer, 2001; Meyer, 2001). When users rely on the warning indicator, they will be more liberal as the probability of signal feedback increases, and more conservative as it decreases. Comparison of empirical and optimal response bias provides an indication of whether an observer over- or underrelies on the automation.

In the CID context, a hit occurs when a user shoots at a hostile soldier. An FA occurs when a user shoots at a friendly soldier. Thus,

$$P(Hit) = P(Shoot \mid Hostile), \text{ and}$$
$$P(FA) = P(Shoot \mid Friendly).$$

When a CID system provides feedback on target identity, it is likely to affect how the user responds. Thus, we are likely to obtain different P(Hit) and P(FA) scores given "unknown" or "friend" feedback. When users receive "unknown" feedback,

$$P(Hit) = P(Shoot \mid \text{"Unknown"} \cap Hostile), \text{ and}$$
$$P(FA) = P(Shoot \mid \text{"Unknown"} \cap Friendly).$$

P(Hit) and P(FA) can be similarly computed for the "friend" feedback situation. However, given that "friend" feedback is extremely unlikely with a hostile target, $P(\text{"Friend"} \cap Hostile)$ is likely to be near zero. Hence we do not compute sensitivity or bias measures in the "friend" feedback context.

A measure of response bias, $\beta_{unknown}$ can be computed from the *P(Hit) and P(FA)* scores in the "unknown" context. If "unknown" feedback indicates that a target is likely to be hostile, a user would likely show a reduced $\beta_{unknown}$ relative to a no-aid situation. The SDT approach allows for a computation of an optimal β. Thus, in the no-aid situation,

$$\beta_{optimal} = \frac{P(Friendly)}{P(Hostile)} \times \frac{C(Miss) + V(Hit)}{C(FA) + V(CR)}.$$

Similarly, $\beta_{optimal}$ can be defined for the "unknown" feedback context:

$$\beta_{optimal\_unknown} = \frac{P(Friendly|''Unknown'')}{P(Hostile|''Unknown'')} \times \frac{C(Miss) + V(Hit)}{C(FA) + V(CR)}.$$

One difficulty in computing optimal response bias—or any reliance method, for that matter—is quantifying the payoff structure. One can avoid this difficulty by computing the response bias difference instead. The difference of response bias is defined as the natural logarithm of the ratio between response bias in the manual condition and the aided condition with "unknown" feedback (Maltz & Shinar, 2003; Murrell, 1977):

$$\ln(\beta_{manual} : \beta_{unknown}) = \ln\beta_{manual} - \ln\beta_{unknown}.$$

The natural logarithm is used to transform the response bias β (whose scale is compressed for liberal values relative to conservative) to a linear scale on which negative values indicate liberal bias and positive values indicate conservative bias. Consequently, the optimal response bias difference can be defined as

$$\ln \beta_{optimal\_manual} - \ln \beta_{optimal\_unknown} = \ln\left[\frac{P(Friendly)}{P(Hostile)} \times \frac{C(Miss)+V(Hit)}{C(FA)+V(CR)}\right] - \ln\left[\frac{P(Friendly|''Unknown'')}{P(Hostile|''Unknown'')} \times \frac{C'(Miss)+V'(Hit)}{C'(FA)+V'(CR)}\right].$$

In the CID context, the factors that are likely to affect the payoff structures (e.g., the rules of engagement) are likely to be the same with or without the presence of an automated CID system. Therefore, for our study, we assume that the decision payoffs were equal in manual and "unknown" feedback conditions, simplifying the expression to

$$\ln \beta_{optimal\_manual} - \ln \beta_{optimal\_unknown} = \ln\left[\frac{P(Friendly)}{P(Hostile)}\right] - \ln\left[\frac{P(Friendly|''Unknown'')}{P(Hostile|''Unknown'')}\right].$$

The response bias difference can also be used to measure the difference in a user's reliance on two aids whose reliability differs.

We chose response bias difference as the measure of participants' reliance on the automation feedback in our experiment for three reasons. First, it defined the optimal reliance level, without which the judgment of the appropriateness of reliance would be difficult. Second, the magnitude of the algebraic response bias difference between conditions is readily interpreted. Third, it allowed us to define optimal reliance without quantifying the decision payoffs.

Similar analytical modeling approaches have been developed by others. For example, the expected value model (Sheridan & Parasuraman, 2000) offers a measure of the economic value of automation versus human performance by discriminating between different decisions types (i.e., hit, false alarm, correct rejection, and miss) and their associated costs and benefits. Sorkin and Woods (1985) developed an SDT modeling approach for a mixed (human and automated) alerting system but assumed, as in many process control situations, that the user did not attend to an event until alerted by the automated system. It is likely that these other approaches, if suitably adapted to the CID context, would yield similar results. Regardless, the application of an analytic approach to the CID context is novel.

In the automation literature, user responses to automation have been said to take two forms, compliance and reliance (Meyer, 2001, 2004). *Compliance* refers to the user's responding as if there is a signal in the world given signal feedback from automation. *Reliance* refers to the user's responding as if there is noise in the world given noise feedback (Dixon & Wickens, 2006; Dixon, Wickens, & McCarley, 2007; Meyer, 2001, 2004). In the CID context, we use the phrases *reliance on "unknown" feedback* and *reliance on "friend" feedback* to refer to participants' compliance with and reliance on the automation, respectively.

Hypotheses

There is a need to establish experimentally if users will adjust to changes in automation reliability of a CID system. There is also a need to specify the appropriateness of that change—to indicate if there is overreliance or underreliance. Finally, the relation between belief, reliance, and trust should be examined.

We sought to examine the effectiveness of using system reliability information to support appropriate trust and reliance on a CID aid. We predicted that disclosure of aid reliability information would allow participants to calibrate their reliance on the aid. As aid reliability changes, participants who were informed of aid reliability should adjust their reliance more appropriately than those who were not informed. We also hypothesized that there should be a correlation between belief and trust in the CID aid and that trust in, and reliance on, the aid should also be correlated.

METHOD

Participants

Recruited were 26 university students with normal or corrected-to-normal visual acuity. Complete data were collected from 24 participants. Each participant was paid Can$30, and a bonus of Can$10 was given to the participant

Figure 1. Screen shots of the IMMERSIVE Simulation.

who had the greatest accuracy in the experimental task.

Apparatus

The experimental simulation was built with modules of the commercial first-person shooter game Unreal Tournament 2004. The simulation was installed on Dell OptiPlex GX270 desktop computers with 20-inch (51 cm) UltraSharp 2000FP flat panel liquid crystal displays.

Figure 1 shows a screenshot of the participants' view and the appearance of targets in the simulation. Participants could control the direction of the weapon and fire on targets in the scene. As is evident in the figure, the targets were distinguished by appearance cues, such as different weapons, color and patterns of uniforms, helmets, and face camouflage.

Experimental Design

The experiment employed a 3 (aid reliability: no aid, 67% reliability, and 80% reliability) × 2 (reliability disclosure: uninformed vs. informed) mixed design with aid reliability manipulated within participants and reliability disclosure manipulated between participants. In the no-aid condition, participants conducted the CID task manually. In the aided conditions, a reliability of x percent meant that when the aid provided "unknown" feedback, it identified a hostile target correctly x percent of the time. Reliability disclosure refers to whether participants were explicitly informed of the reliability of the "unknown" feedback. In each condition, half the targets were friendly and the other half were hostile.

Measures

Belief. To measure belief about aid reliability, participants in the uninformed group were asked to estimate the failure rate of the "unknown" feedback after each aided mission block.

Trust. Participants' trust was measured by questionnaire after each aided mission block (see Table 1). The mean score of the first 11 items was used to assess general trust in the aid (extracted from Jian, Bisantz, & Drury, 2000). Two further items assessed trust in the "unknown" and "friend" feedback.

Reliance. Participants' reliance on the "unknown" feedback was analyzed with the response bias difference approach.

Performance. Participants' error rate and response time (RT) were used to measure general CID performance. Error rate assessed participants' response in all trials for each experimental condition. RT was defined as the elapsed time between the target appearance and the first shot at the target. Thus, RT was recorded only when a participant shot at a target.

Task Procedure

The experiment comprised a training session and three mission blocks. The training session familiarized participants with the simulation and the appearance of friendly and hostile targets. During the training, participants were first shown pictures of friendly and hostile targets and were guided to look for the differences in the appearance of the two target types. Then participants were asked to judge the identity of targets and shoot at hostile targets in 60 training trials. The experimenter provided feedback

TABLE 1: Items Used to Measure Trust in the Combat Identification System

1. The aid is deceptive.
2. The aid behaves in an underhanded (concealed) manner.
3. I am suspicious of the aid's outputs.
4. I am wary of the aid.
5. The aid's action will have a harmful or injurious outcome.
6. I am confident in the aid.
7. The aid provides security.
8. The aid is dependable.
9. The aid is reliable.
10. I can trust the aid.
11. I am familiar with the aid.
12. I can trust that *blue* lights indicate Canadian soldiers.
13. I can trust that *red* lights indicate terrorists.

about whether the target was friendly or hostile after each training trial. Participants then completed three mission blocks: one no-aid block and two aided blocks with aid reliabilities of 67% and 80%, respectively. Block order was counterbalanced across participants. Each block contained 120 trials and lasted about 20 min.

On each trial, either a friendly or a hostile soldier appeared. The experimental task was to identify the target in the scene and shoot as soon as possible if it was an enemy. In the no-aid condition, several cues were available from the target appearance. In the two aided conditions, there was one additional cue available: the feedback from the CID aid. Therefore, when using a CID aid, participants could combine the feedback with the appearance cues to make the final judgment. After a soldier was killed or a trial ended, participants were prompted to indicate the confidence level of their engagement decision, from 5 = *highly confident* to 1 = *not at all confident*. Each participant was informed of the hostile and friendly target probability.

Participants were advised that they would have a CID aid to assist them in two of the three blocks. The aid responded when the weapon was pointed at a target. When it identified a friendly soldier, a blue light was shown on the weapon ("friend" feedback); otherwise, a red light was shown ("unknown" feedback). All participants were told that whereas the "friend" feedback was always correct, the "unknown" feedback was fallible. All participants were required to pass a short test to make sure that they were aware of the different reliabilities of the "friend" and "unknown" feedback. Participants in the informed group were told the failure rate of "unknown" feedback at the beginning of each aided block. At the end of an aided block, participants filled out the trust questionnaire, and those in the uninformed group estimated the failure rate of the "unknown" feedback.

RESULTS

To increase normality and stabilize variances, an arcsine transformation was applied to all probability data used in ANOVA (Howell, 1992; Winer, Brown, & Michaels, 1991). Effect size, r, was calculated for significant effects.

Performance

A 3 (aid reliability: no aid, 67%, 80%) × 2 (reliability disclosure: uninformed, informed) ANOVA was conducted on transformed error rate. A main effect of aid reliability was obtained, $F(2, 44) = 9.70$, $p < .01$. Planned contrasts revealed significantly less error in the 80% reliability condition ($M = .12$) than in the no-aid condition ($M = .20$), $F(1, 22) = 15.11$, $p = .01$, $r = .64$. There was no significant difference between the 67% reliability condition ($M = .16$) and the no-aid condition, $F(1, 22) = 3.40$, $p = .08$, $r = .37$. Neither the effect of reliability disclosure nor the interaction was significant, $F < 1$ in both cases.

A mean RT was computed for each participant in each condition and submitted to a 3 × 2 ANOVA. No significant effect was obtained.

Reliance

Reliance on the "unknown" feedback. For each aid reliability condition, the empirically determined z-transformed Receiver Operating Characteristics (ROCs) (Macmillan & Creelman, 1991) were calculated for each participant (except for 5 participants who did not make a miss error in at least one of the three mission blocks). A least squares regression was used to find a slope for each ROC. The two-tailed one-sample t test revealed that there was no significant difference between the empirical slopes and the null value 1.00, $t(54) = 1.30$,

p = .20, M = 1.20, SE = 0.151. Therefore, the equal-variance assumption of SDT was not violated in this study. This result indicated that d' and β were appropriate indices of detection sensitivity and response bias, respectively.

We examined the difference in participants' response bias between the no-aid and 67% reliability conditions (i.e., no aid – 67%), between the 67% and 80% reliability conditions (i.e., 67% – 80%), and between the no-aid and 80% reliability conditions (i.e., no aid – 80%). Because the third difference condition was dependent on the first two, we included only the first two difference conditions in the ANOVA. However, we tested all three difference conditions in separate t-tests when comparing the empirical reliance to the optimal level.

A 2 (difference condition: no aid – 67%, 67% – 80%) × 2 (reliability disclosure: uninformed, informed) ANOVA on lnβ difference revealed a main effect of reliability disclosure, $F(1, 22)$ = 5.11, p = .03, r = .43. The participants in the informed group changed their response bias more dramatically than those in the uninformed group. The main effect of difference condition was not significant, nor was the interaction, both $F < 1$.

As shown in Figure 2, the difference in response bias for the uninformed group was less than optimal in all three conditions, $t(11)$ = –2.57, –2.89, and –6.60; p = .03, .02, .00; r = .61, .65, .89; for no aid – 67%, 67% – 80%, and no aid – 80%, respectively. In contrast, the adjustment of response bias for the informed group was not significantly different from the optimal value (lnβ[opt] = .693, shown by the lower dashed line in Figure 2) from the no-aid condition to the 67% reliability condition, $t(11)$ = –0.93, p = .55, or from the 67% reliability condition to the 80% reliability condition, $t(11)$ = –1.32, p = .22. However, their adjustment was smaller than the optimal value (lnβ[opt] = 1.386, the higher dashed line in Figure 2) from the no-aid condition to the 80% reliability condition, $t(11)$ = –2.36, p = .04, r = .58. Therefore, in the 67% reliability condition, whereas the informed group relied on the "unknown" feedback appropriately, the uninformed group did not rely on the feedback enough. In addition, the uninformed group also did not adjust their reliance enough to accommodate the change of aid reliability from 67% to 80%.

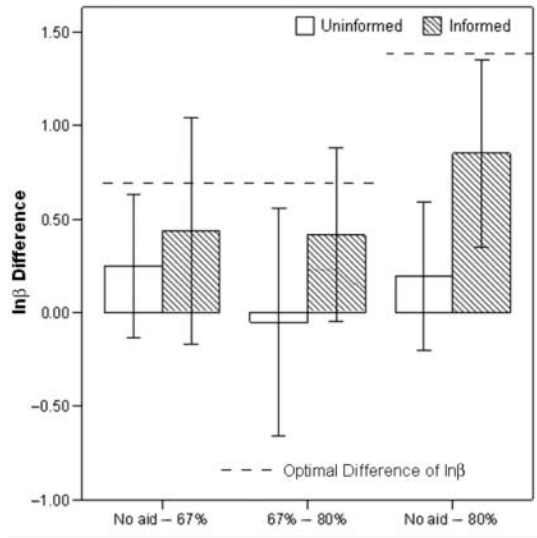

Figure 2. Comparison of participants' difference of response bias and the optimal difference.

Reliance on the "friend" feedback. Response bias difference could not be calculated on the "friend" feedback trials because hostile targets did not appear in these trials. However, because "friend" feedback was 100% correct, if participants relied on it, fewer errors would occur. Thus, we used error rate as an indirect measure of reliance on "friend" feedback. Two Wilcoxon signed rank tests were conducted to examine the effect of aid reliability on error rate in the "friend" feedback trials for each reliability disclosure group. Similarly, two Mann-Whitney tests were conducted to examine the effect of reliability disclosure on the error rate in the "friend" feedback trials in each aid reliability condition. The informed group had greater error rates on the "friend" feedback trials than the uninformed group regardless of reliability condition: For the 67% reliability condition, M = .07 vs. .04, U = –36.50, p = .03, r = –.61; for the 80% reliability condition, M = .06 vs. .04, U = –39.00, p = .05, r = –.57. There was no significant effect for aid reliability in either reliability disclosure condition.

Belief, Trust, and Reliance

The trust scores for the whole aid, the trust scores for "unknown" feedback, the belief scores (uninformed group only), and reliance

TABLE 2: Correlation Among Belief, Trust, and Reliance

Categories of Measures			Belief	Trust		Reliance
			Estimate	Whole Aid	"Unknown"	lnβ Difference
Belief	Estimates	τ(24)	—	–.27*	–.50**	–.11
Trust	Whole aid	τ(48)	—	—	.19*	.11
	"Unknown"	τ(48)	—	—	—	.21*
Reliance	lnβ difference		—	—	—	—

Note. *$p < .05$. **$p < .01$.

scores (lnβ difference) were calculated for each participant in each aided condition. The scores were submitted to a set of one-tailed Kendall's τ correlations to examine the relationships among belief in, trust in, and reliance on the CID aid (see Table 2). Because the informed group was not asked about beliefs, the correlations between belief and trust and between belief and reliance were calculated only for the uninformed group in the two aided conditions.

As shown in Table 2, participants' trust in the whole aid and trust in the "unknown" feedback were both negatively correlated with their estimate of the "unknown" feedback failure rate. This indicated that the more reliable the "unknown" feedback was thought to be, the more participants trusted the whole aid and the "unknown" feedback. However, there was no significant correlation between belief and reliance. The correlation between trust and reliance was calculated using both groups' data in the two aided conditions. Participants' response bias difference had a positive relationship with their trust in the "unknown" feedback but not trust in the whole aid. This result suggested that participants' reliance on "unknown" feedback was more closely related to specific trust in the "unknown" feedback than to general trust in the whole aid.

DISCUSSION

Performance

Previous studies have not shown an improvement in overall accuracy with imperfect CID systems (Dzindolet et al., 2000; Dzindolet, Pierce, Beck, et al., 2001; Dzindolet, Pierce, Pomranky, et al., 2001; Karsh et al., 1995; Kogler, 2003). Our results showed a benefit for an 80% reliable CID system but did not show a benefit for a 67% reliable system. In an informal meta-analysis, Wickens and Dixon (2007) showed that lower levels of imperfect automation did not improve performance relative to a manual condition, but higher levels did, with a cutoff around 70% (see also Parasuraman, Sheridan, & Wickens, 2000). Many previous CID studies used low reliability levels, such as 60% (Dzindolet, Pierce, Beck, et al., 2001; Kogler, 2003). Our results seem to align well with the results of Wickens and Dixon (2007) and suggest that CID performance is indeed similar to the results typically obtained in other automation domains.

Another factor that may explain the observed benefit of the CID system in this study was the low cost of reliance. The cost of reliance on automation can increase the likelihood of automation disuse. For example, in the study by Karsh et al. (1995), there was a 0.75-second wait for aid feedback, but participants were required to respond as quickly as possible. Disuse of the automation was observed. In contrast, in our study, there was no delay after the weapon was aimed at a target.

Appropriateness of Reliance

The appropriateness of reliance depended on reliability disclosure. The uninformed group's response bias difference was generally less than optimal. In contrast, the informed group's response bias difference did not differ significantly from the optimal value in the no aid – 67% or the 67% – 80% condition. Thus, the informed group seemed to resolve the changing reliability of "unknown" feedback better than the uninformed group. Nonetheless, the adjustment of the informed group's response bias was still less than optimal in the no aid – 80% condition.

In general, these results serve as another example of the "sluggish beta" phenomenon, in which the human response to a change in target probability is often less than the ideal magnitude (Chi & Drury, 1998; Wickens & Hollands, 2000).

In an unexpected result, instruction of aid reliability did not guarantee complete reliance on the 100% reliable "friend" feedback. In fact, the informed group had slightly higher error rates than the uninformed group. Examinations of trust, self-confidence, and reliance scores; scrutiny of potential outlier influences; and evaluations of participants' explanations all failed to produce a systematic explanation for the result. Beck, Dzindolet, and Pierce (2007) used appraisal error and intent error to explain the different causes of suboptimal automation usage. Appraisal errors are caused by incomplete knowledge of the relative utilities of the available options, whereas intent errors refer to the case when the users are aware of the utilities but intentionally disregard this information. In this study, given that participants were aware that the "friend" feedback was 100% reliable, their noncompliance with "friend" feedback may have represented intent errors.

This study used a difference in response bias between two conditions to examine reliance on automation. We believe this is novel in the CID context. Given its definition of optimal reliance level and independence from decision payoff structures, response bias difference has clear advantages over measures of general consistency, difference between misuse and disuse, or even response bias alone.

Belief, Trust, and Reliance

Participants' belief in and reliance on "unknown" feedback were not correlated with each other. However, they both correlated with trust in "unknown" feedback. These results support the hypothesis that trust in an aid mediates the relationship between a user's belief about an aid's capabilities and reliance on the aid (Lee & See, 2004). If the causal relationships between belief and trust and between trust and reliance exist, the differences between the informed and uninformed groups' reliance could be seen as a causal chain starting from belief about aid reliability. Whereas the informed group's more accurate belief about the aid reliability was directly supported by instructions, the uninformed group needed to form belief based on interaction with the aid. This discrepancy between the two groups' direct knowledge about the "unknown" feedback failure rate was reflected in the difference in reliance on the "unknown" feedback.

CONCLUSION

In this study, compared with the manual condition, the CID accuracy was improved in the 80% reliability condition but not in the 67% reliability condition. This supports the previous finding (Parasuraman et al., 2000) that high-reliability automation is, in general, a prerequisite for improved performance in automated tasks.

The findings suggest that trust in the "unknown" feedback from CID systems mediates the relationship between the belief in, and reliance on, the "unknown" feedback. Therefore, disclosing the aid reliability level to users appeared to positively influence the appropriateness of trust, which in turn contributed to the more appropriate reliance on CID systems that we observed. These findings may assist with the design of information displays for CID systems, augmenting "friend" or "unknown" feedback with reliability information. The findings may also be useful for training soldiers to assess system reliability individually. Soldiers can trust and rely on CID systems better if they are alert to contextual factors that affect system reliability, such as the friend:foe ratio, the presence of civilians, the battlefield terrain, and possible signal interference.

We identified and briefly critiqued two approaches to measuring reliance on CID systems. These approaches consider reliance from the perspective of general consistency, and the difference between misuse and disuse. Although these approaches offer some insight into the reliance construct, we favor the response bias difference approach for two reasons. First, it offers a definition of the appropriate reliance level. Second, it clearly illustrates the difference in participants' reliance between different aided conditions. However, when applying the response bias difference approach to other settings, researchers should exercise caution

in examining the assumption of equal decision payoff structures between the manual condition and aided conditions.

We note that relative to a real combat situation, the experimental task was simplified in terms of workload, stress levels, and the variation in appearance within each force. Furthermore, the participants were not soldiers. Nonetheless, the study does show that imperfect automation can improve CID performance and that human observers will adjust reliance more appropriately when informed of the reliability levels. Testing in a more realistic setting with soldiers as participants should be used to validate these findings.

ACKNOWLEDGMENTS

This research was conducted at the University of Toronto under Contract No. W7711-06800/001/TOR from Defence Research and Development Canada–Toronto. We thank Francois Bernier of Defence Research and Development Canada–Valcartier, who provided us access to the IMMERSIVE simulator used to run the simulations for our experiment. We also thank Liuba Mamonova for her assistance in conducting this study.

At least one of the authors of this article is an employee of the Canadian government and created the article within the scope of their employment. As a work of the Canadian government, the content of the article is in the public domain.

REFERENCES

Beck, H. P., Dzindolet, M. T., Pierce, L. G. (2007). Automation usage decisions: Controlling intent and appraisal errors in a target detection task. *Human Factors*, *49*, 429–437.

Biros, D. P., Daly, M., & Gunsch, G. (2004). The influence of task load and automation trust on deception detection. *Group Decision and Negotiation*, *13*, 173–189.

Bisantz, A. M., & Pritchett, A. R. (2003). Measuring judgment interaction with displays and automation: A lens model analysis of collision detection. *Human Factors*, *45*, 266–280.

Boyd, C. S., Collyer, R. S., Skinner, D. J., Smeaton, A. E., Wilson, S. A., Krause, D. W., et al. (2005). Characterization of combat identification technologies. In *IEEE International Region 10 Conference* (pp. 568–573). Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments* (2nd ed.). Berkeley: University of California Press.

Chi, C., & Drury, C. G. (1998). Do people choose an optimal response criterion in an inspection task? *Institute of Industrial Engineers Transactions*, *30*, 257–266.

Cohen, M. S., Parasuraman, R., & Freeman, J. T. (1998). Trust in decision aids: A model and its training implications. In *Proceedings 1998 Command and Control Research and Technology Symposium* (pp. 1–37). Washington, DC: Department of Defense C4ISR Cooperative Research Program.

Dixon, S. R., & Wickens, C. D. (2006). Automation reliability in unmanned aerial vehicle control: A reliance-compliance model of automation dependence in high workload. *Human Factors*, *48*, 474–486.

Dixon, S. R., Wickens, C. D., & McCarley, J. S. (2007). On the independence of compliance and reliance: Are automation false alarms worse than misses? *Human Factors*, *49*, 564–572.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., Dawe, L. A., & Anderson, B. W. (2000). Misuse of an automated decision making system. In *Conference on Human Interaction With Complex Systems 2000* (pp. 81–85). Urbana-Champaign, IL: Beckman Institute for Advanced Science and Technology.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., Dawe, L. A., & Anderson, B. W. (2001). Predicting misuse and disuse of combat identification systems. *Military Psychology*, *13*, 147–164.

Dzindolet, M. T., Pierce, L., Pomranky, R., Peterson, S., & Beck, H. (2001). Automation reliance on a combat identification system. In *Proceedings of the 45th Annual Meeting of the Human Factors and Ergonomics Society* (pp. 532–536). Santa Monica, CA: Human Factors and Ergonomics Society.

Friend or Foe identification system. (2007). FEMSWISS AG. Retrieved May 4, 2009, from http://www.army-technology.com/downloads/whitepapers/training/file218/

Gimble, T. F., Ugone, M., Meling, J. E., Snider, J. D., & Lippolis, S. J. (2001). *Acquisition of the battlefield combat identification system* (Report No. D-2001-093). Washington, DC: U.S. Department of Defense, Office of the Inspector General.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics.* New York: Wiley.

Howell, D. C. (1992). *Statistical methods for psychology* (3rd ed.). Belmont, CA: Duxbury.

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.

Jones, C. (1998, January/February). The battlefield combat identification system: A Task Force XXI response to the problem of direct fire fratricide. *ARMOR*, pp. 43–46.

Karsh, R., Walrath, J. D., Swoboda, J. C., & Pillalamarri, K. (1995). *Effect of battlefield combat identification system information on target identification time and errors in a simulated tank engagement task* (Technical Report ARL-TR-854). Aberdeen Proving Ground, MD: U.S. Army Research Laboratory.

Kogler, T. M. (2003). *The effects of degraded vision and automatic combat identification reliability on infantry friendly fire engagements.* Unpublished master's thesis, Virginia Polytechnic Institute and State University, Blacksburg, VA.

Lee, J. D., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics, 35*, 1243-1270.

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

Lerch, F. J., Prietula, M. J., & Kulik, C. T. (1997). The Turing effect: The nature of trust in expert systems advice. In P. J. Feltovich, K. M. Ford, & R. R. Hoffman (Eds.), *Expertise in context: Human and machine* (pp. 417–448). Cambridge, MA: MIT Press.

Lowe, C. (2007). *Cutting through the fog of war.* Retrieved April 15, 2009, from http://www.defensetech.org/archives/003496.html

Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide.* New York: Cambridge University Press.

Madhavan, P., & Wiegmann, D. A. (2004). A new look at the dynamics of human-automation trust: Is trust in humans comparable to trust in machines? In *Proceedings of the Human Factors and Ergonomics Society 48th Annual Meeting* (pp. 581–585). Santa Monica, CA: Human Factors and Ergonomics Society.

Maltz, M., & Meyer, J. (2001). Use of warnings in an attentionally demanding detection task. *Human Factors*, *43*, 217–226.

Maltz, M., & Shinar, D. (2003). New alternative methods of analyzing human behavior in cued target acquisition. *Human Factors*, *45*, 281–296.

Masalonis, A., & Parasuraman, R. (2003). Effects of situation-specific reliability on trust and on usage of automated air traffic control decision aids. In *Proceedings of the Human Factors and Ergonomics Society 47th Annual Meeting* (pp. 533–537). Santa Monica, CA: Human Factors and Ergonomics Society.

Meyer, J. (2001). Effects of warning validity and proximity on responses to warnings. *Human Factors*, *43*, 563–572.

Meyer, J. (2004). Conceptual issues in the study of dynamic hazard warnings. *Human Factors*, *46*, 196–204.

Muir, B. M., & Moray, N. (1996). Trust in automation: 2. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, *39*, 429–460.

Murrell, G. A. (1977). Combination of evidence in a probabilistic visual search and detection task. *Organizational Behavior and Human Performance*, *18*, 3–18.

Parasuraman, R., & Mouloua, M. (1996). Monitoring of automated systems. In R. Parasuraman & M. Mouloua (Eds.), *Automation and human performance: Theory and applications* (pp. 91–115). Mahwah, NJ: Lawrence Erlbaum.

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.

Sheridan, T. B., & Parasuraman, R. (2000). Human versus automation in responding to failures: An expected-value analysis. *Human Factors*, *42*, 403–407.

Sheridan, T. B., & Parasuraman, R. (2006). Human–automation interaction. In R. S. Nickerson, (Ed.), *Reviews of human factors and ergonomics* (Vol. 1, pp. 89–129). Santa Monica, CA: Human Factors and Ergonomics Society.

Sherman, K. (2000). Combat identification system for the dismounted soldier. In *Proceedings of SPIE 2000: Digitization of the Battlespace V and Battlefield Biomedical Technologies II* (pp.135–146). Bellingham, WA: the International Society for Optical Engineering.

Sherman, K. B. (2002). Combat ID coming for individual soldiers. *Journal of Electronic Defense*, *25*, 34–35.

Snook, S. A. (2002). *Friendly fire: The accidental shootdown of U. S. Black Hawks over northern Iraq.* Princeton, NJ: Princeton University Press.

Sorkin, R. D., & Woods, D. D. (1985). Systems with human monitors: A signal detection analysis. *Human–Computer Interaction*, *1*, 49–75.

St. John, M., Smallman, H. S., Manes, D. I., Feher, B. A., & Morrison, J. G. (2005). Heuristic automation for decluttering tactical displays. *Human Factors*, *47*, 509–525.

Wang, L., Jamieson, G. A., & Hollands, J. G. (2008). Selecting methods for the analysis of reliance in automation. In *Proceedings of the 52nd Annual Meeting of the Human Factors and Ergonomics Society* (pp. 287–291). Santa Monica, CA: Human Factors and Ergonomics Society.

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

Wickens, C. D., & Hollands, J. G. (2000). *Engineering psychology and human performance* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Winer, B. J., Brown, D. R., Michels, K. M. (1991). *Statistical principles in experimental design* (3rd ed.). New York: McGraw-Hill.

Lu Wang is a PhD student in the Department of Systems and Information Engineering at the University of Virginia in Charlottesville. She obtained her MASc in mechanical and industrial engineering from the University of Toronto in 2008.

Greg A. Jamieson is an associate professor in the Department of Mechanical and Industrial Engineering at the University of Toronto. He obtained his PhD in mechanical and industrial engineering from the University of Toronto in 2003.

Justin G. Hollands is head of the Human Systems Integration Section at Defence Research and Development Canada–Toronto and an adjunct associate professor in the Department of Mechanical and Industrial Engineering at the University of Toronto. He obtained his PhD in psychology from the University of Toronto in 1993.

Date received: July 30, 2008
Date accepted: April 30, 2009