Developing Human-Machine Interfaces to Support Appropriate Trust and Reliance on Automated Combat Identification Systems

Progress Report for Milestones 1, 2 & 3

September 17, 2007

PW&GSC Contract Number: W7711-068000/001/TOR

Prepared by:

Greg A. Jamieson
Project Lead

Lu Wang
Research Assistant

ABSTRACT

This research tested the effects of system reliability information and interface features on human trust and reliance on individual combat ID systems. Experiment I showed that participants had difficulty in estimating the reliability of the 'unknown' feedback from these systems. Providing the reliability information led to appropriate reliance on that feedback. Experiment II showed that participants' trust in the 'unknown' feedback was influenced by the system's activation mode and the 'unknown' feedback form, but their reliance on 'unknown' feedback was not affected. In addition, a new method was proposed to measure reliance on automation. This measure was used effectively in both experiments, and demonstrated several advantages over previous methods. Finally, implications for the design of interfaces for individual combat ID systems and the training of infantry soldiers were drawn from the results of the studies.

EXECUTIVE SUMMARY

This report summarizes work completed to date on the project "Developing Human-Machine Interfaces to Support Appropriate Trust and Reliance on Automated Combat Identification Systems". As specified in the Statement of Work, the major tasks for the three milestones were:

- Milestone 1: Experiment planning
- Milestone 2: Experiment I execution
- Milestone 3: Experiment II execution

For Milestone 1, we started by reviewing the academic literature pertaining to operator trust in automation and then proposed two experiments to address gaps in this literature. Experiment I was designed to determine whether providing system reliability information would lead to appropriate trust and reliance on combat identification (ID) systems. Experiment II was designed to test whether the differences between simulated combat ID systems and real system prototypes would influence operator trust in, and reliance on, the system feedback. In addition, because previous empirical studies have not clearly defined reliance on automation, they lack a criterion for appropriate reliance. To address this problem, we have developed a new experimental method for assessing operator reliance on automation.

For Milestone 2, we conducted Experiment I and analyzed the data. The findings suggest that participants' beliefs about the system reliability and their trust in the system feedback are positively correlated. The findings further indicate that participants' trust in the feedback is positively correlated with their reliance on the feedback. The participants had difficulty in estimating the system reliability. Informing them of the aid reliability information led to appropriate trust in, and reliance on, the system.

For Milestone 3, we conducted Experiment II and analyzed the data. The findings suggest that, although the participants' reliance on the system feedback was generally not affected by activation mode and feedback form, they had a clear preference for those interface features and their trust in the system feedback was influenced by them. These findings indicate that the dissimilarity between the simulated aids used in previous studies and the real system prototypes can lead to changes in humans' trust in the system feedback. This change in trust attitude may then influence their intention to use the system. Experiment II also suggests that designers of a human machine interface (HMI) for combat ID systems should carefully consider both the content and the form of feedback provided to operators.

In the next stage of this project, we will design an HMI for combat ID systems based on the results of Experiments I and II. The design process will follow a generic User-Centered Design methodology, including a specification of users, an analysis of user tasks, and an iterative design process.

MILESTONE 1: EXPERIMENT PLANNING

During the Experiment Planning phase of this project, the following work was completed:

- Review of the literatures pertaining to fratricide, combat ID systems, human-automation interaction, and empirical studies related to combat ID systems.
- Design of Experiment I and Experiment II,
- Preparation of experimental materials (see Appendix A – J),
- Obtaining of approval for the two experiments from the Social Sciences and Humanities Research Ethics Board at the University of Toronto (U of T) (see Appendix K),
- Modification of a combat ID virtual simulation to enable the conduct of the two experiments, and
- Conducting of pilot studies for Experiment I and Experiment II to test the combat ID simulation and the methodology of the two experiments.

This section first reviews literatures related to this project. It then presents the motivation, hypothesis and design of Experiment I and Experiment II. The last part of this section describes the modifications made to the combat ID simulation for this project.

Literature Review

Problem Statement and Purpose of Research

Friendly fire has caused heavy casualties in the contemporary warfare. One of the primarily reasons was soldiers' deficiency in distinguishing friends and enemies in the chaotic combat zone (Gimble, Ugone, Meling, Snider, & Lippolis, 2001; Jones, 1998). With the purpose of improving the soldiers' combat ID ability, a variety of technical solutions have been developed. The fact that the contemporary warfare involves many dismounted urban operations draws attention to one of these technologies – the individual combat ID system (ICIDS) (Lowe, 2007). The drawback of such a system is that it cannot positively identify any solider without a working transponder device (K. Sherman, 2000; K. B. Sherman, 2002; "SIMLAS," 2006). Therefore, when the signal is not positive, the solider can be hostile, neutral or friendly. Soldiers seem to have problems relying on this imperfect automation appropriately, which calls in question the benefit of this technology (Briggs & Goldberg, 1995).

Humans are prone to misuse and disuse imperfect automation (Parasuraman, Molloy, & Singh, 1993). Previous studies have consistently demonstrated that humans' trust in automation is a major factor that determines their reliance on the automation (Lee & See, 2004). The goal for the current project is to test factors that affect the humans' trust and reliance on the combat ID systems, and thereby helping them better utilize the systems and reducing fratricide incidents.

Fratricide

Fratricide, as a military term, is defined by U.S. Army Training and Doctrine Command (TRADOC) as "the employment of friendly weapons and munitions with the intent to kill the enemy or destroy his equipment or facilities, which results in unforeseen and unintentional deaths or injury to friendly personnel"(U.S. Department of Army, 1993, p.1). It is also commonly referred to as friendly fire. Heavy casualties from fratricide and collateral damage

have been an 'inconvenient truth' throughout the history of war (Hughes, 1996; Shrader, 1982). The statistics show that fratricide accounted for at least 10% of the total U.S. casualties in World War II, Viet Nam War, the first Persian Gulf War, as well as other major wars in the 20th century (Shrader, 1982; Steinweg, 1995). The U.S. Marine Corps have admitted 23 fratricide incidents (82 casualties) since 2001 ("Frightening Friendly," 2007). The errant lethal shooting of one Canadian solider in August 2006 ("Canadian Killed," 2006), the U.S. aircraft strafing of a Canadian platoon in September 2006 ("Soldier Killed," 2006), and the killing of eight Iraqi policemen in February 2007 are just a few of the latest tragedies ("US Air," 2007). These fratricide incidents not only negatively impact troop morale and tactical effectiveness, they also induce public recrimination and have a devastating effect on the family members of victims (Snook, 2002; Wilson, Salas, & Priest, 2007; Young, 2005). These dramatic costs highlight the imperative need to discover the causes of fratricide and develop possible countermeasures (Bourn, 2002; Rierson & Ahrens, 2006)

Fratricide is usually caused by a combination of many factors: improper tactics, inadequate group communication and technical problems are examples (Frisconalti, 2005; Snook, 2002; Wilson et al., 2007). However, human errors in combat ID clearly played a major part in most incidents (Jones, 1998; Regan, 1995). Combat ID is "the process of attaining an accurate characterization of entities in a combatant's area of responsibility to the extent that high-confidence, real-time application of tactical options and weapon resources can occur" ("Defense Science," 2000, p.1). Soldiers' ability to perform combat ID task can greatly affect the rate of fratricide incidents. However, soldiers seem to be incapable of conducting combat ID task effectively in the 'fog of war' – the extremely chaotic battlefield. Historical data manifests that ground units are even less capable at effectively conducting combat ID. Specifically, during all the wars in the 20th century, about 46% of fratricide incidents occurred in situations solely involving ground units (Bourn, 2002).

Individual Combat Identification System

Over 25 technologies have been proposed and/or developed to help soldiers to identify friends and foes in the battlefield, such as the radio frequency ID tag, battlefield target ID system, individual combat ID system, and blue force tracking system (Boyd et al., 2005). Among these technologies, the individual combat ID system is intended for infantry soldiers to identify other friendly infantry soldiers. Several prototypes of the individual combat ID system have been developed and evaluated in the past few years (K. Sherman, 2000; K. B. Sherman, 2002; "SIMLAS", 2006).

The individual combat ID system is a cooperative technology which consists of an interrogator and a transponder (Boyd et al., 2005) (see Figure 1). A soldier who is equipped with the interrogator sends a directional encrypted laser query to an unidentified soldier by pressing the activation button and aiming his/her weapon at the unidentified soldier. If the interrogated soldier is wearing an appropriate transponder, the transponder will decode and validate the interrogation message, and send a coded radio frequency (RF) reply. If the interrogator receives a correct RF reply, the light-emitting diode (LED) on his weapon will blink to give a 'friend' feedback. Otherwise, based on the available information of the system prototypes, no explicit feedback will be given to the interrogating solider (K. Sherman, 2000; "SIMLAS," 2006). The battlefield target ID system which is intended for combat vehicles also operates through a similar query and response process. However, when it does not receive a correct reply, it will send out an 'unknown' reply (Gimble et al., 2001). For example, the systems installed on the M2A2 Bradley Infantry Fighting Vehicle will signal a flashing red light for friendly targets and a constant

yellow light for unknown targets (Jones, 1998). The 'no feedback' from the individual combat ID systems after failing to receive correct reply can be seen as an implicit 'unknown' reply.

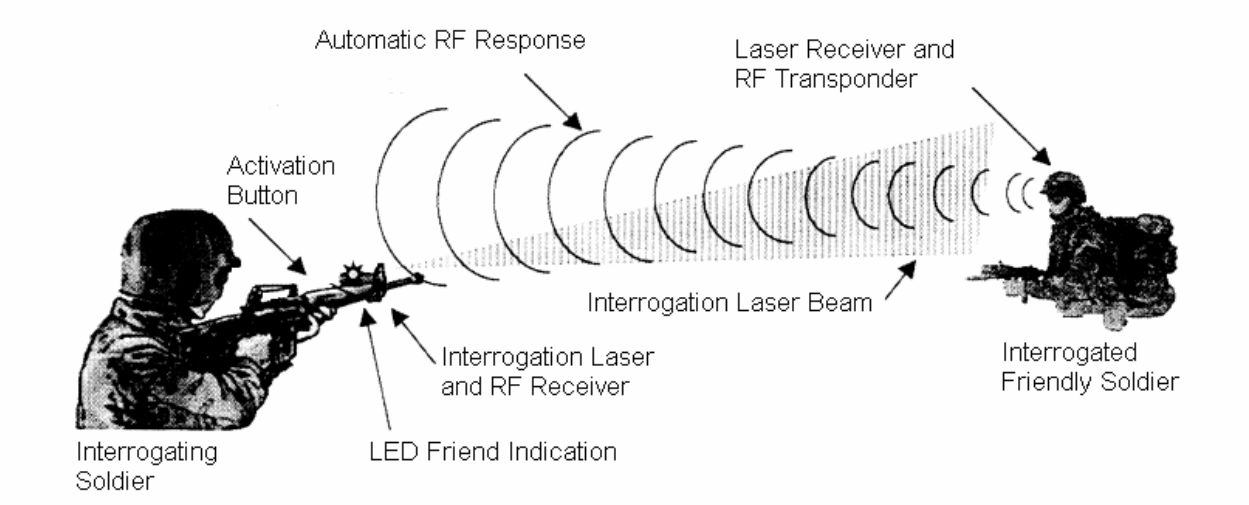


Figure 1. The working mechanism of the individual combat ID system

(adapted from K. Sherman, 2000, p. 137)

The reason why the non-positive feedback is 'unknown' instead of 'enemy' is that, these interrogation/response combat ID systems cannot positively identify enemies. Other than a target being hostile, many possibilities exist when the normal interrogation/response process has not been completed. For instance, the target could be a civilian, from some neutral force, friendly but lacking a proper transponder, or friendly but equipped with a proper transponder that cannot be recognized due to the electronic signal garbling in the combat zone, a dead battery in a transponder or the incorrect selection of system mode (Snook, 2002).

Therefore, while the 'friend' response corresponds to friendly forces[1], the 'unknown' response may not correctly indicate hostile forces (Boyd et al., 2005). Some developers claimed that the system can correctly identify 97.5% of the friendly targets when the targets were within 1100 meters (K. Sherman, 2000). However, these tests were conducted in a controlled environment. In real battlefield, many factors can interfere with the communication between the interrogator and transponder. For example, terrain can sometimes block the line-of-sight of the radio wave, and result in system failures (Boyd et al., 2005; Snook, 2002). Hence, the success rate may not be so high.

This leads to the question, what is the probability that a target is hostile given that the feedback from a combat ID aid is unknown? This probability is called the reliability of 'unknown' feedback in this project. The reliability of 'unknown' feedback is contingent on the percentage of hostile forces in the battlefield and the percentage of non-hostile forces that a system can positively recognize. In order to help soldiers correctly interpret the 'unknown' feedback, they should be aware of these two pieces of information. The percentage of recognizable non-hostile forces is influenced by many contextual factors, such as the locations of detected targets, neutral forces and civilians in the battlefields. The percentage of enemies also varies with the changes in

[1] In a few situations, it is possible that an interrogator designates a hostile soldier as friendly. For example, if a properly equipped friendly soldier is very close to the hostile soldier. However, the chance of misidentification of friendly force is very small, so we did not consider this type of failure in the scope of this study.

battlefield. Therefore, it is important to find some way to deliver the moment-by-moment information to soldiers.

Human-Automation Interaction

The feedback from the combat ID systems is provided to soldiers to inform their combat ID decision. However, since these systems are not perfectly reliable, they cannot replace soldiers' judgment based on their own visual examination and situational awareness. This technology limitation poses new questions to the system designers and implementers: can the soldiers rely appropriately on the feedback from these imperfect combat ID systems? If they cannot, what should be done to help them to adjust their reliance to the optimal level?

Human-automation interaction has been studied for many years due to the prevalence of automation related accidents and incidents (Bainbridge, 1983; Lee & See, 2004; Parasuraman & Riley, 1997). In this chapter, the literature pertaining to this subject is reviewed, and the means to support good human-automation interaction is discussed.

Problematic Use of Automation

It is a common misconception that automation is introduced to replace human operators with the purpose of alleviating human errors (Sheridan, 1996). Yet, operators are still required to monitor and supervise most automated systems (Mosier & Skitka, 1996; Sheridan, 2002). This new job is not error-free and sometimes is even more demanding than their original manual work. Engineers usually design and implement automation without the consideration of its impact on operators, such as their mental workload, manual skill and situational awareness (Bainbridge, 1983; Parasuraman, Sheridan, & Wickens, 2000; Skitka, Mosier, & Burdick, 2000). As a result, many serious accidents and incidents happened because operators failed to use the automated systems properly (Lee, 2006; Parasuraman & Riley, 1997).

Parasuraman and Riley (1997) claimed that the human' problematic use of automation primarily falls into two categories: disuse and misuse of automation. Disuse of automation refers to the case where operators fail to rely on reliable automation (e.g. Dzindolet, Peterson, Pomranky, Pierce, & Beck, 2003; Karsh, Walrath, Swoboda, & Pillalamarri, 1995). The benefit of automation cannot be obtained if it is disused. Misuse of automation occurs when operators overly rely on unreliable automation (e.g. Bagheri & Jamieson, 2004; Parasuraman et al., 1993; Skitka, Mosier, & Burdick, 1999). The operators who misuse automation often fail to intervene when the automation malfunctions. Clearly, the problematic uses of automation are closely related to operators' decision to rely or not to rely on automation (Riley, 1996)

Trust and Reliance

The operators' reliance decision is affected by a number of factors such as self-confidence, perceived risk, trust in automation, time constraints, and fatigue (Dzindolet, Pierce, Beck, & Dawe, 1999; Lee & Moray, 1992; Mosier & Skitka, 1996; Riley, 1994; Riley, 1996). Among these factors, trust in automation has been examined in many empirical studies and is deemed as a critical factor that influences humans' reliance on automation (Lee &See, 2004; Lerch, Prietula, & Kulik, 1997; Madhavan & Douglas, 2004; Masalonis & Parasuraman, 2003; Muir, 1987; Muir 1989).

Trust is defined as "the attitude that an agent will help achieve an individual's goal in a situation characterized by uncertainty and vulnerability" (Lee & See, 2004, p. 54). Trust in automation can

be an outcome of operators' belief about automation characteristics, and a cause of their reliance of the automation (Lee & See, 2004; Sheridan & Parasuraman, 2006). Appropriate trust is desirable, because trust determines, in part, the operators' strategies to use the automation (Lee & Moray, 1992; Muir & Moray, 1996).

The appropriateness of trust is the match between the operators' trust and the true capability of the automation. It can be described from three perspectives, calibration, resolution and specificity. Calibration is "the correspondence between a person's trust in the automation and the automation's capabilities" (Lee & See, 2004, p. 55). Resolution describes "how precisely a judgment of trust differentiates levels of automation capability" (Lee & See, 2004, p. 55). Specificity is "the degree to which trust is associated with a particular component or aspect of the trustee" (Lee & See, 2004, p. 56). Specificity can be both functional and temporal. Functional specificity indicates the specificity of trust to different subfunctions or modes of a system. Temporal specificity indicates the sensitivity of trust to the changing automation capabilities in different context (Lee & See, 2004).

Support of Appropriate Trust

In order to support appropriate operator trust in automation, it is critical to understand the basis of trust and the processes guiding the evolution of trust. High calibration, resolution and specificity of trust can only be achieved when the information concerning the basis of trust is made available to operators in a way that is consistent with these trust evolution processes (Lee & See, 2004).

The basis of trust is comprised of two elements: the focus of trust and the information supporting trust (Lee & See, 2004). 'The focus of trust' is the entity to be trusted. It can be described according to the level of detail. Sometimes it is general – the trust may focus on a whole complex system; sometimes it is specific – the trust may focus on a particular subfunction or a mode. 'The information supporting trust' is what informs the beliefs about 'the focus of trust'. It can be classified into three categories (Lee & Moray, 1992; Lee & See, 2004). The first is the performance information. It describes the current and past operation of the automation and its attributes. The second is the process information. It describes the algorithms and processes that are underlying the automation behaviors. The third is the purpose information. It describes the intention for which the automation was originally designed. The availability of these three types of information to the operators is critical in forming a correct belief about the automation capability. Because the beliefs about the automation capabilities greatly affect operators' trust in automation, to generate appropriate trust in an automated system, all these three types of information for each level of detail should be provided to operators (Lee & See, 2004; Sheridan & Parasuraman, 2006).

Human operators interpret 'the information supporting trust' through three different cognitive processes: analytic process, analogical process, and affective process (Lee & See, 2004). In the analytic process, trust is an outcome of the rational analysis based on objective evidences. An example is Cohen, Parasuraman, and Freeman's (1998) normative trust model – the Argument-based Probabilistic Trust (APT) model. The analytic process is very cognitively demanding, especially when the automation is sophisticated. A less cognitively demanding process is the analogical trust, which is dependant on intermediaries (e.g. reputation and gossip) and category judgment that associate trust with automation characteristics and working context. For instance, analogical trust can be inferred from computer etiquette. Parasuraman & Miller (2004) tested the effect of the automation communication style (i.e. non-interruptive/patient vs.

interruptive/impatient) on participants' trust in a high-criticality automated system. They found that trust had a positive relationship with the computer etiquette. The least cognitively demanding process that governs trust is the affective process. Affective trust is the emotional response that people feel about the automation. In the context of website interaction, researchers found that the website interface exercises a great influence on the users' feeling of the credibility, ease of use, and risk of a website, which ultimately can lead to the changes of their trust in the website (Corritore, Kracher, & Wiedenbeck, 2003; Kim & Moon, 1998; Milne & Boza, 1999).

Reliance on Combat Identification Systems

This chapter reviews the field investigations and empirical studies about the reliance on combat ID systems (e.g. Briggs & Goldberg, 1995). Since the measure of the reliance on combat ID systems is deficient in the previous empirical studies (Dzindolet, Pierce, Beck, Dawe, & Anderson, 2000, 2001a; Dzindolet, Pierce, Pomranky, Peterson, & Beck, 2001b; Kogler, 2003), this project proposes a new method to measure the reliance on these systems.

Influence of Combat Environment

Field investigations in the combat ID domain suggest that certain factors in the intense combat environment can strongly influence soldiers' use of these systems and decision criterions (Briggs & Goldberg, 1995). One factor is the extreme cost associated with an incorrect recognition. Fearing the penalties, some Gulf War soldiers chose to rely on themselves and turned off their combat ID aids (Dzindolet et al., 2000). In addition, soldiers have a strong bias to identify a target as a foe given any doubt in a tactical situation (Briggs & Goldberg, 1995). Another factor is the extreme time pressure experienced on the battlefield. The longer a soldier takes to make a decision, the more dangerous the surrounding conditions become. When a soldier is not able to conduct visual identification effectively and the combat ID system is not reliable, the soldier should hold off on making identification decision until more information is available. However, stressful situations sometimes result in soldiers making immature engagement decisions that may trigger friendly fire incidents (Frisconalti, 2005; Steinweg, 1995; Young, 2005).

Measure of the Reliance on Combat ID Systems

A few empirical studies have been conducted to explore humans' reliance on the combat ID systems (Dzindolet et al., 2000, 2001a, 2001b; Karsh et al., 1995; Kogler, 2003). One limitation of the previous studies is the lack of a clear definition of 'appropriate reliance' on the combat ID aid. Some researchers (e.g., Kogler, 2003) compared the participants' performance when they had feedback from a combat ID aid (i.e. aided condition) with their performance when they did not have aid feedback (i.e. no aid condition). And when the participants in the no aid condition outperformed those in the aided condition, concluded that the participants did not rely on the aid appropriately. There are two shortcomings to this approach. First, because two types of mistakes can be made (i.e., friendly fire or missing a foe), important information is lost when these two mistakes are combined together. Second, performance improvement does not always correspond to appropriate reliance. For example, if a soldier's manual accuracy is 50% and he is given a 100% reliable combat ID aid. If the soldier's combat ID accuracy improves from 50% to 70% after the he gets that aid. Would we say that he/she relies on the aid appropriately? The answer is no, because the soldier's performance can be further improved if he relies more on the aid.

Dzindolet et al. (2001a) posed an alternative approach to judge the appropriateness of reliance (i.e. the 'misuse and disuse' approach). They operationally defined misuse (over-reliance on automation) as the participants' error rate in the trials that the aid feedback was wrong, and disuse (underutilization of automation) as the error rate in the trials that the aid feedback was

correct. The researchers asserted that if the misuse rate is larger than the disuse rate, then the participants tend to rely on the aids, and consequently "misuse was more likely than disuse" (Dzindolet et al., 2001a, p12). The limitation of this approach is that the authors jumped directly from reliance to misuse (i.e., over-reliance). This conclusion is premature because even though participants relied on the aid, they could rely on it at an appropriate level and thus not 'misuse' it. Take an extreme case for example. If an aid is highly reliable, say the aid's error rate is 1%; meanwhile participants' manual performance is inferior, say the participants' error rate is 50%. If participants rely completely on the aid, the misuse rate will be 100% and the disuse rate will be 0%. According to the 'misuse and disuse' approach, because misuse rate is larger than disuse rate, the participants are likely to misuse the aid. However, this conclusion is inaccurate because absolute reliance in this case is appropriate since the participants themselves are completely incompetent for the task.

To overcome the deficiency of the two previous approaches, this research proposes an innovative way to measure reliance on automation (i.e. 'response bias' approach). This new approach is based on Signal Detection Theory (SDT) (Macmillan & Creelman, 1991; Wickens & Hollands, 2000). In SDT, the participants' performance is characterized by two performance indicators—sensitivity and response bias. When a soldier receives an aid feedback, what changes should be his/her expectation of the probability that the target is friendly or hostile. This change, according to SDT, should influence the setting of the response bias but not of the sensitivity. If this premise is true, then the reliance on the aid can be measured based on the change of the response bias. Since the reliability of the 'unknown' and 'friend' response is different, the appropriate reliance on these two types of feedback is different. The reliance should be measured separately for the situations when the soldiers receive 'unknown' feedback and the situations when they receive 'friend' feedback.

The participants' response when the aid gave 'unknown' feedback can be mapped onto the following outcome matrix of SDT (see Table 1).

Table 1. The outcome matrix in the condition that the aid gave 'unknown' feedback

		States of the World	
		P(Enemy|Unknown)	**P(Friend|Unknown)**
Participant Response	**Shoot**	Hit (H) P(H|Unknown) Value=V(H)	False Alarm (FA) P(FA|Unknown) Cost=C(FA)
	Not Shoot	Miss (M) P(M|Unknown) Cost=C(M)	Correct Rejection (CR) P(CR|Unknown) Value=V(CR)

Since the 'unknown' feedback indicates a target is likely to be hostile, the more the solider relies on it, the more liberal the soldier should become. If their response bias becomes more liberal, then the solider shows reliance on the 'unknown' feedback; if the response bias becomes more conservative, then the soldier tends to reject the 'unknown' feedback; if the response bias does not change, then the soldier is ignorant of the 'unknown' feedback. Furthermore, the appropriateness of reliance can be described by the match between a soldier's response bias and the optimal response bias. With a soldier's sensitivity being constant, the closer their response bias is to the optimal value, the fewer mistakes he/she will make. According to SDT,

$\beta_{optimal} = \frac{P(Friend \mid Unknown)}{P(Enemy \mid Unknown)} \times \frac{V(CR) + C(FA)}{V(H) + C(M)}$. The first part of the formula depends on the reliability of the 'unknown' feedback, and the second part of the formula depends on the payoffs of different decision outcomes.

Similarly, when a solider receives a 'friend' feedback, the more conservative the solider becomes, the more he/she relies on the 'friend' feedback. When the 'friend' feedback is always correct, the appropriate reliance is complete compliance with it. When the 'friend' feedback is fallible, then, $\beta_{optimal} = \frac{P(Friend \mid 'Friend' \ feedback)}{P(Enemy \mid 'Friend' \ feedback)} \times \frac{V(CR) + C(FA)}{V(H) + C(M)}$.

The 'response bias' approach is superior to the previous methods because not only does it show whether the soldiers rely on the aid or not, it also clearly identifies the level of reliance that will lead to the best performance. This approach is expected to lead to more informative interpretation of the results from future empirical studies. In both experiments of this research, the 'response bias' approach was used to analyze the participants' reliance on the 'unknown' feedback.

Implications from Empirical Studies

This section reviews the empirical studies that are closely related to humans' use of combat ID systems (Dzindolet et al., 2000, 2001a, 2001b; Karsh et al., 1995; Kogler, 2003).The first part of this section describes the methods and the findings of each of these empirical studies. These details would be referred back later in discussing the similar and different results between this research and these previous studies. The second part of this section focuses on discussing the support of participants' appropriate trust in combat ID aid in these studies.

Summary of Empirical Studies

Karsh et al. (1995) tested the influence of a combat ID system on the soldiers' accuracy and speed in a simulated tank engagement task. The control group conducted the task manually. The treatment group was given the combat ID aid and could get the automation feedback 0.75 seconds after they interrogated the targets. The 'friend' feedback was a red light and the 'unknown' feedback was a yellow light. The treatment group was informed that the probability that the aid correctly recognized friendly targets was set to be 90% and the probability that the aid mistakenly identified hostile targets as friendly was set to be 4%. During the experiment, half the targets were friendly, the other half were hostile. The results indicated that the treatment group only interrogated the targets in about 14.5% of all the trials. No significant differences were found in the accuracy and speed of the identification decision between the control group and the treatment group.

Dzindolet et al. (2000, 2001a, 2001b) performed two studies to understand humans' interaction with a combat ID system. In the first study (Dzindolet et al., 2000, 2001a), the task was to detect whether there was a soldier in terrain slides. A solider was contained in 24% of the slides. The simulated combat ID aid offered two types of feedback. They would be shown on the screen after the appearance of the slides. One feedback was the word 'PRESENT' and a red circle; the other feedback was the word 'ABSENT' and a green circle. The two types of feedback were both fallible. The accuracy was 60%, 75% or 90% respectively in three different groups (aided groups). The participants were informed of the aid reliability. In addition, there is a controlled group who worked without the aid (unaided group). The results indicated that there was no

performance difference among the three aided groups and the unaided group. Based on the 'misuse and disuse' reliance measure method, they concluded the participants over-relied on the aid feedback regardless of the aid reliability.

In the second study (Dzindolet et al., 2001b), in addition to the solider detection task, the participants also had a secondary task which was to respond to audio stimuli. The first group of participants conducted the tasks manually (unaided group), and the second and third groups were assisted by an automated aid in the soldier detection task (aided groups). Both aided groups went through 200 trials before the formal test to get experience of the aid, but only the third group was explicitly informed of the aid reliability. The automation was always correct when it provided 'ABSENT' response, but was only correct 67% of the time when it responded 'PRESENT'. A solider was contained in 50% of the slides. Two measures of reliance behavior were taken in this study. First was the percentage of trials in which participants asked to view the slides again. Second was the reliance on aid feedback ('misuse and disuse' method). Compared with the unaided group, the two aided group requested fewer second views when the aid responded with the 100% correct feedback 'ABSENT', but similar number of second views when the aid responded with the fallible feedback 'PRESENT'. Hence, the first measure (second views) indicated that the two aided groups reacted reasonably to the aid feedback. First, they seemed to be aware of that only the 'ABSENT' feedback was more reliable in the two types of feedback. Second, they seemed to be cautious of the fallible feedback 'PRESENT' in that they requested as many second views as the unaided group when the aid responded 'PRESENT'. However, the second measure (reliance on aid) revealed that the participants did not rely on the aid appropriately. Although the third group was explicitly informed of the aid reliability, their disuse of the perfectly reliable 'ABSENT' feedback was not significantly less than the fallible 'PRESENT' feedback. The second group even showed greater disuse of the 'ABSENT' feedback than the 'PRESENT' feedback. The inappropriate reliance strategy of the two aided groups led them to make more errors than the unaided group.

Kogler (2003) examined the effects of the degraded vision and the reliability of a combat ID system on soldiers' target identification performance. The three levels of degraded vision were 75%, 10% and 2% transmissivity. There were three levels of aid reliability. In the no aid condition, the participants conducted the task manually; in the low reliability aid condition, the aid identified 60% of the friendly targets; in the high reliability aid condition, the aid correctly identified all of the friendly targets. Never would the aid mistakenly identify a hostile target as friendly. Among the 15 targets, 5 were friendly and the others were hostile. The aid provided feedback through the participants' headsets, it would be either 'friend-friend-friend' or 'unknown'. All of the participants were explicitly informed with the aid reliability in all the test conditions. The results indicated that the degraded vision led to slow response and more identification errors. The participants were almost unable to identify the targets in the 2% transmissivity level by themselves. On average, they engaged 4.5 out of 5 friendly targets. For the two other transmissivity levels, the average number of friend fire engagement was about 1.8 in the no aid conditions. While the 100% reliable aid completely eliminated friendly fire engagement in all three transmissivity conditions, the 60% reliable aid did not significantly reduce the number of friendly fire engagement in the 10% and 75% transmissivity conditions, but significantly reduced this number in the 2% transmissivity condition. In addition, the number of missed threat targets was not significantly different among all the transmissivity and reliability conditions.

Providing Information about Basis of Trust

The benefit or harm of combat ID systems on the overall performance depends on whether the participants can rely on the automation feedback appropriately. When fallible feedback is relied on properly, an imperfect system can improve task performance. For instance, St. John and Manes (2000) found that participants in a target detection task successfully used unreliable automation to direct and facilitate their search. However, in the empirical studies about combat ID systems, sometimes the participants disused a reliable aid (Karsh et al., 1995), sometimes they overly relied on the feedback from unreliable aids (Dzindolet et al., 2000, 2001a, 2001b). Overall, the results from the reviewed empirical studies suggest that the performance was not improved by the combat ID systems unless the system reliability was 100% or the manual performance was very deficient (Kogler, 2003). To make it even worse, providing an aid sometimes even deteriorated the performance (Dzindolet et al., 2001b).

Researchers speculated that the suboptimal use of the aid in their experiment might be caused by the inappropriate trust in the aid (Dzindolet et al., 2001a). Since trust was not recorded in these studies (Dzindolet et al., 2000, 2001a, 2001b; Karsh et al., 1995; Kogler, 2003), it is impossible to determine whether the improper reliance was caused by inappropriate trust or not. But it is worth looking back to their experiment setting to see whether the participants got enough information to form the appropriate trust in the combat ID systems? The trust in automation is based on three types of information about the automation: performance, process and purpose (Lee & See, 2004). Table 0-2 summarizes the availability of these three types of information in the previous studies. Overall, the participants seemed to get all the information or at least most of the information required to generate appropriate trust. Therefore, even when the participants were conscious of the system characteristics, they did not rely on it appropriately.

Table 0-2. The availability of trust basis information in the previous studies

Study	Types of Information		
	Performance	Process	Purpose
Karsh et al., 1995	Yes (instruction)	No	Yes (instruction)
Dzindolet et al., 2000, 2001a	Yes (instruction)	Yes (instruction)	Yes (instruction)
Dzindolet et al., 2001b	Yes (instruction or experience)	Yes (instruction)	Yes (instruction)
Kogler, 2003	Yes (instruction)	No	Yes (instruction)

These results are unexpected. Humans' reliance on automation is affected by their trust in automation (Masalonis & Parasuraman, 2003; Muir & Moray, 1996); and the appropriateness of their trust is strongly affected by the correspondence between their perception of the system capability and its actual capability (Cohen et al., 1998; Dzindolet et al., 2000; Lee & See, 2004). Therefore, providing information of the aid capability should benefit the appropriate trust and reliance.

There are two possible explanations for this unexpected result. First, although the necessary information was available, their trust was still not appropriate. Trust in automation is guided by the analytical, analogical and affective processes (Lee & See, 2004). The instruction about the aid characteristics might only guide the analytical aspect of trust. It is hard to anticipate if there were other factors in those experiments that might influence the analogical and affective aspects

of trust. Second, the participants' trust was appropriate, but their reliance on the combat ID systems was not appropriate. Trust is not the only determinant of reliance (Parasuraman & Mouloua, 1996; Lee & See, 2004). Other factors, such as self-confidence, workload and time constraints, may also influence reliance behavior. Therefore, even the participants trusted the aids appropriately, they might not rely on the aids appropriately.

To order to determine whether information about aid capability can lead to appropriate trust in the aid feedback and to find out whether trust is related to humans' suboptimal reliance on the combat ID system, the first experiment of this project was conducted.

Experiment I

Objective

Previous studies suggest that participants tended to rely inappropriately on the combat ID aid even if they were informed of the aid capability (Dzindolet et al., 2000, 2001a, 2001b; Karsh, et al., 1995). The two primary objectives of this experiment are: first, to examine effectiveness of using aid reliability information to support appropriate trust and reliance on the aid; second, to scrutinize the relationships among the participants' belief about the aid reliability, their trust in the aid, and their reliance on the aid. The secondary purpose of this experiment is to test the feasibility of using response bias as an indication of participants' reliance on the combat ID aid.

Hypotheses

Three hypotheses are derived from the literatures reviewed in the previous section. Assertions are followed by support from the literatures.

Hypothesis 1: There is a positively correlation between the participants' belief of the capability of a combat ID aid and their trust in the aid, and between their trust in the aid and their reliance on the aid.

Hypothesis 2: When working with the aid, the participants who are informed of the aid reliability will trust and rely on the aid more appropriately than those who are not informed.

Humans' "trust (in automation) can be both a cause and an effect" (Sheridan & Parasuraman, 2006, p. 100). As a cause, it influences humans' use of automation (Masalonis & Parasuraman, 2003; Muir & Moray, 1996); as an effect, it is dependent on their belief about the aid capability (Lee & See, 2004). Trust is more likely to be appropriate when the information related to the automation capability is available (Cohen et al., 1998; Dzindolet et al., 2000). These conclusions from the previous literatures provide grounds for Hypothesis 1 and 2.

Hypothesis 3: The fallible 'unknown' feedback would change participants' response bias but not their sensitivity in the combat ID task.

The 'unknown' feedback might change the participants' expectation of the probability that a potential target is hostile. According to Signal Detection Theory (SDT), humans' response bias will vary with their expectation of the target probability, whereas their sensitivity will stay constant (Macmillan & Creelman, 1991; Wickens & Hollands, 2000). If this hypothesis holds, it will give support to a new method to measure reliance on combat ID system (i.e. 'response bias' reliance measure method).

Experimental Design

A 3×2 mixed design was employed. The within-subjects factor was the aid reliability which had 3 levels: no aid, 67% and 80%. In the no aid condition, the participants did not get the aid and they conducted the combat ID task manually. The reliability of x percent means that when the aid sends out an 'unknown' feedback, x percent of the time it correctly identifies a terrorist target. The between-subjects factor was the instruction of the aid reliability – whether or not the participants were informed of the reliability of the 'unknown' feedback. Since in the no aid condition, the participants performed the task without the aid, the instructions were identical for the two groups of participants when they were in the no aid condition.

The experiment was comprised of three mission blocks with different aid reliabilities. The order of conditions was counterbalanced separately across all of the participants. Each block consisted of 120 trials and only one target appeared in each trial. For each block, the targets in half of the trials were friendly and the other half were hostile.

Experiment II

Objective

The results from the former studies about combat ID aid (Dzindolet et al., 2000, 2001a, 2001b; Karsh et al., 1995; Kogler, 2003) may contribute to helping infantry soldiers better rely on the combat ID systems. However, the simulated combat ID systems in these studies were dissimilar to real system prototypes in the activation method and the indication of 'unknown' feedback (Sherman, 2000; "SIMLAS", 2006). In some previous studies, the simulated systems responded automatically after the appearance of the stimuli (Dzindolet et al., 2000, 2001a, 2001b). In addition, the 'unknown' feedback was always explicit, such as, a yellow light (Dzindolet et al., 2000, 2001a, 2001b; Karsh et al., 1995; Kogler, 2003). However, the individual combat ID systems are not prototyped to work like this (Sherman, 2000; "SIMLAS", 2006). First, to interrogate a target, the soldiers need to manually activate the aid. Second, the interrogator does not send out explicit signal when no reply is received.

Will the conclusions from these studies hold if the ecological validity of the experiment is improved by making the simulated system more similar to the real systems? This experiment is the initial attempt to test whether the interface features of a combat ID system, such as the activation method and the indication of 'unknown' feedback, would cause the participants to react differently.

Hypotheses

Humans' trust in automation is influenced by their perceptions of the credibility, ease of use, and risk of the automation (Corritore et al., 2003). The content and format of an automation interface can affect these perceptions, even though they do not necessarily reflect the true capabilities of the automation (Lee & See, 2004). In this experiment, two interface features, activation mode and feedback form, were manipulated. Therefore, the participants' perception of the aid and trust in the aid might vary in different conditions. However, it is hard to predict the specific effects of these two features. Take the effect of the activation mode for example. The auto mode requires less work in activation, but it also makes the participants lose control of the aid. It is difficult to determine whether the participants would deem it easier to use than the manual mode or not. Therefore, this experiment is exploratory and no specific hypothesis was proposed.

Experimental Design

A 2×2 repeated-measures design was employed. The first factor introduced two levels of system activation modes: automatic (auto) and manual. In the auto mode, the system was always turned on and it responded automatically whenever the weapon was pointed at a target as in Experiment I; in the manual mode, the system was off unless the participants pressed an activation button. The second factor introduced two forms of 'unknown' feedback: red light and no light. For the 'red light' condition, the aid responded a red light to hostile targets, just liked the aid in Experiment I. However, for the 'no light' condition, it did not send out any response when it considered a target a terrorist.

This experiment consisted of four blocks with different combination of activation modes and feedback forms. The automation reliability was constant at the 67% reliability level. The order of conditions was counterbalanced separately across all the participants. Each block consisted of 60 trials with one target appearing in each trial. For each block, the targets in half of the trials were friendly and the other half were hostile.

Modification of Combat Identification Simulation

The synthetic task environment used in this experiment was IMMERSIVE (Instrumented Military Modeling Engine for Research using Simulation and Virtual Environments). It was developed by Defence Research and Development Canada at Valcartier, utilizing the modules of a commercial first-person shooter game – Unreal Tournament 2004. Figure 2 shows the screenshot of a simulated scene.

Figure 2. Participant's view in IMMERSIVE

In IMMERSIVE, experimenters can create scenarios by setting terrains, combat activities as well as characteristics of forces. Friendly and hostile forces are distinguished by differences in uniforms, weapons, actions, and feedback from the combat ID systems (See Figure 3).

Figure 3. Different uniforms and weapons for friendly and hostile forces

The simulation was installed on two Dell OptiPlex GX270 desktop computers in the Cognitive Engineering Laboratory at U of T. The technical specifications of these two computers were the same: Intel Pentium 4 800Mhz FSB processor, 80GB 7200ROM Parallel ATA, NVIDIA GeForce 6800 Graphics Card, SoundMAX Integrated Digital Audio, and FPS 1500 speakers. The 20-in UltraSharp 2000FP flat panel monitors were set at High Color (32bit) resolution, 800 x 600 pixels.

Seven students from U of T were recruited for the pilot study of Experiment I and four students from U of T were recruited for the pilot study of Experiment II. Based on the data and feedback from these two pilot studies, we modified several aspects of IMMERSIVE to enable it to meet the requirement of Experiment I and Experiment II:

- Map
  - Modified file: UT2004\CombatIDMod\Maps\TOR-CCID_Experiments.ut2
  - Modification:
    - We adjusted the illumination level by changing the brightness and radii of the lights in the participants' view and the sun light. This change was made to control the difficulty of the combat ID task.
    - We added four paths that the simulated targets (soldiers) would follow in these two experiments. In addition, the targets' movement was changed from walking to running to increase the time pressure of the combat ID task. Several visual blocks, such as fences, were added to the map to ensure that the difficulty of the combat ID task were approximately identical for the four paths.
- Cuing interface
  - Modified file: UT2004\CombatIDMod\CombatIDInterface\classes\ CIDCuingInterface.uc
  - Modification: The question "Rate the probability that the target is hostile" was changed to "How confident are you that you have made the correct decision". And the original six points scale (0%, 20%, 40%, 60%, 80%, 100%) was changed to a five points scale (1 – not at all confident, 2 – slightly confident, 3 – somewhat confident, 4 – confident, 5 – highly confident). This change was made based a former combat ID empirical study (Dzindolet et al., 2001a).
- Appearance time of the cuing interface
  - Modified file: UT2004\CombatIDMod\CombatIDEngine\Classes\CIDTargetsManager.uc
  - Modification: Originally, the cuing interface could only be set to appear at a predetermined time. The modification associated the appearance time with the participants' action. After the modification, if the participants killed a target, the cuing interface would pop up right away; otherwise it would pop up at the predetermined time. This modification decreased the participants' waiting time for the cuing interface after they made their engagement decision.
- Message
  - Modified file: UT2004\CombatIDMod\CombatIDMessage\Classes\CIDExperimentationFinishMessage.uc

- Modification: The message at the end of each block was changed from “The experiment is over” to “The mission block is over”. This change was made because each experiment in this project consisted of several mission blocks.

MILESTONE 2: EXPERIMENT I EXECUTION

This section presents an overview of the work conducted during the Experiment I Execution phase of this project. The first part of this section describes the data collection; the second part presents the statistical analysis of the collected data. The last part discusses these results and compares them to the findings in the previous studies (Dzindolet et al., 2000, 2001a, 2001b; Karsh et al., 1995; Kogler, 2003).

Data Collection

Participants

26 students with normal visual acuity from U of T were recruited. Complete data were collected from 24 participants and only those data were used for analysis. Half the participants were informed of the aid reliability and half of them were not. Each participant was paid $ 30 CAD for his/her participation, and a bonus $ 10 CAD was given to the top performer who had the greatest accuracy in engagement decisions. A similar compensation scheme was used by previous studies and has been found to be adequately motivating (e.g. Dixon & Wickens, 2006).

Tasks and Procedures

The experiment took approximately two and a half hours to complete. To be qualified to take part in this experiment, each participant was required to pass a vision test. Their visual acuity was measured with a Snellen eye chart, and their ocular dominance was measured using the Porta Test (Roth et al., 2002). If the participants passed the visual test, they would be given the informed consent form (See Appendix A). After signing the informed consent form, they filled out a short demographic information survey (See Appendix B). The participants were then given a sheet of instructions about the experiment procedure. At the end of the instruction session, the participants completed a questionnaire to demonstrate their comprehension of the instructions (see Appendix C).

In the instructions, the participants were asked to imagine they were in a battlefield. They were told that their primary task was to identify targets in the scene and shoot them if they believed they were enemies. Their final score was determined by the accuracy and speed of their engagement decision. The participants were told that they were going to complete 3 mission blocks, each consisting 120 trials. The identities of targets were randomized with the constraints that half of the trials were friends and the other half were enemies. In each trial, a target appeared in the scene for approximately 10 seconds. The participants were instructed to make an engagement decision as soon as possible. The decision could be either ‘shoot’, or ‘do not shoot’. After a target was killed or a trial ended, the participants would be asked to indicate the confidence level of their decision on a five-point Likert scale ranging from not at all confident (1) to highly confident (5) (Dzindolet et al., 2001b).

The participants were advised that they would have an aid to assist them in 2 of the 3 blocks. When the aid identified a friendly solider, it would respond a ‘friend’ feedback – a blue light, and otherwise it would respond an ‘unknown’ feedback - a red light. The participants were told that the ‘unknown’ feedback was set to be less than 100% reliable to mimic system failures. It was possible that a red light could be shown when a target was actually friendly. However, the blue light would never appear when a target was hostile. At the end of these two aided mission

blocks, the participants were asked to fill out a questionnaire. For the participants who were informed of the aid reliability, the questionnaire was to measure their trust in the automation. For the participants who were not informed of this information, the questionnaire was to measure their trust in the automation and their estimate of the 'unknown' feedback failure rate (see Appendix D).

After the instructions, the experimenter showed the participants the pictures of friendly and hostile targets. The participants then went through a training session (60 trials) to familiarize themselves with the synthetic task environment and the task. During the training, the experimenter also guided them to improve their identification skills. After that, the participants started the three mission blocks. All of the participants were informed of whether they would have the combat ID aid in the following mission block, but only the participants in the informed group would be told the failure rate of 'unknown' feedback in that mission block. For a complete set of instructions please refer to Appendix C.

Measures and Instruments

Target Identification Performance

Four objective measures were taken in this experiment to examine the combat ID performance:

- False Alarm (FA) rate (friendly fire engagement): the percentage of trials that a participant decides to shoot at a target when it is actually a friendly soldier.
- Miss rate (missed threat targets): the percentage of trials that a participant holds fire on a target when it is actually a terrorist.
- Response time (RT): the elapsed time between when a target appears on the scene and the first shot is fired. Note that the RT was recorded only for those trials that the participants shot at a target.

Misuse and Disuse

Dzindolet et al. (2000) proposed that P(Misuse) and P(Disuse) could be used to indicate participants' reliance on automation. In the context of this experiment, P(Misuse) and P(Disuse) could be defined as:

- P(Misuse): the error rate in the trials that the system sends out 'unknown' feedback and the target is friendly. This error rate is actually the participant's false alarm rate in the trials the aid responds a red light.
- P(Disuse):

  a. the error rate in the trials that the system sends out 'unknown' feedback and the target is an enemy. This is the participant's miss rate in the trials the aid responds a red light.

  b. the error rate in the trials that the system sends out 'friend' feedback. This is participant's false alarm rate in the trials the aid respond a blue light.

SDT Statistics

The participants' reliance on the aid was also measured using response bias. In SDT, there are several ways to express the response bias, such as B, D, C and ß. Among all of the alternative measures, C has the simplest statistical properties (Macmillan & Creelman, 1991, p273), and it was also the measure used in Dzindolet et al.'s study (2001a). Thus, C was used in the analysis

of variance (ANOVA) part of the data analysis. ß, on the other hand, has the advantage that it can be easily compared with the optimal ß calculated based on the target probability and payoffs, thus ß was used to calibrate the appropriateness of reliance. In this experiment, the participants were told that their final score will be determined by the number of the trials that they held fire on a friendly target or shot at a hostile target. Therefore, the value of correct identification of friend was not differentiated from the value of correct identification of terrorist, and there was no cost for wrong decisions. Thus, $\frac{V(CR)+C(FA)}{V(H)+C(M)}=\frac{V(CR)+0}{V(H)+0}=1$. And the optimal ß values are:

for the 67% reliability condition, $\beta_{optimal}=\frac{P(Friend\,|\,Unknown)}{P(Terrorist\,|\,Unknown)}=\frac{33\%}{67\%}=0.50$;

for the 80% reliability condition, $\beta_{optimal}=\frac{P(Friend\,|\,Unknown)}{P(Terrorist\,|\,Unknown)}=\frac{20\%}{80\%}=0.25$.

The calculation of the SDT statistics depends on participants' receiver operator characteristics (ROCs). When the slope of the ROCs in standard coordinates is equal to 1.0, the standard deviations of the noise and signal-plus-noise distributions are equivalent (Dzindolet, et al., 2001a). Then the statistics of participants' sensitivity and response bias are defined by the formulas below:

$$d' = Z_{Hit} - Z_{FA}$$

$$C = -\frac{1}{2}\left[Z_{Hit} + Z_{FA}\right]$$

$$\beta = \exp\left\{d' \times C\right\}$$

If the slope does not equal to 1.0, the sensitivity and response bias are measured by other statistics. Therefore, in the data analysis section, the participants' ROCs were first empirically determined, and then their sensitivity and response bias were calculated using the appropriate statistics.

Subjective Measures

In addition to the objective measures, several subjective measures were taken using a questionnaire (see Appendix D). Questions 1 to 11 were extracted from the first and only empirically determined scale of trust in automation (Jian, Bisantz, & Drury, 2000). Some minor changes were made to original scale based on the pilot study. These 11 questions were used to measure the participants' trust in the whole system. Questions 12 and 13 required the participants to rate their trust in the 'unknown' and 'friend' feedback, respectively. Questions 1 to 13 all used 7 points scales. Question 14 was for the participants in the uninformed group only. They were asked to estimate the failure rate of the 'unknown' feedback.

Data Analysis

The mixed design ANOVA is the primary analysis. Conclusions are made when the effects reach the significance level of .05. Effect sizes are calculated for the contrasts and effects that compared only two levels, $r=\sqrt{\frac{F(1,df_R)}{F(1,df_R)+df_R}}$ (Field, 2005, p. 514). To increase the normality

of the probability data, an arcsine transformation was applied to all the probability data: Transformed Probability Data= 2 * arcsine [Probability Data]$^{1/2}$ (Dzindolet et al., 2001a; Howell, 1992; Winer, 1991). When the assumption of the normality was violated for a measure and the non-normality cannot be corrected by data transformation, non-parametric tests were used instead of the mixed design ANOVA. The Wilcoxon Signed-Rank test was used to examine the effect of the within-subjects factor 'aid reliability' on each group. The Mann-Whitney test was used to examine the effect of the between-subjects factor 'group' in each aid reliability condition. Effect size for non-parametric tests was calculated based on $r = \frac{z}{N}$, z is the z-score of a test, N is the number of observation (Field, 2005, p. 532).

Target Identification Performance

False Alarm (Friendly Fire)

The 3 (aid reliability: no aid, 67%, 80%) X 2 (group: uninformed, informed) ANOVA on the transformed P(FA) revealed a significant main effect of aid reliability, F(1.506[2],44)=10.752, p=.001.Contrasts were performed comparing the no aid condition with the mean of the two aided conditions, and comparing the 67% reliability condition with the 80% reliability condition. There was a significant difference between the no aid condition and two aided conditions, F(1,22)= 9.858, p=.005, r=.556, and a significant difference between the 67% reliability condition and the 80% reliability condition, F(1,22)=13.950, p=.001, r=.623. As seen in Figure 4, the participants made fewer false alarm errors when they had the combat ID aid; and the more reliable the aid was, the fewer false alarm errors they committed. The effect of the group was found to be non-significant, F(1,22)=1.610, p=.218, r=.261, as was the aid reliability X group interaction, F<1.

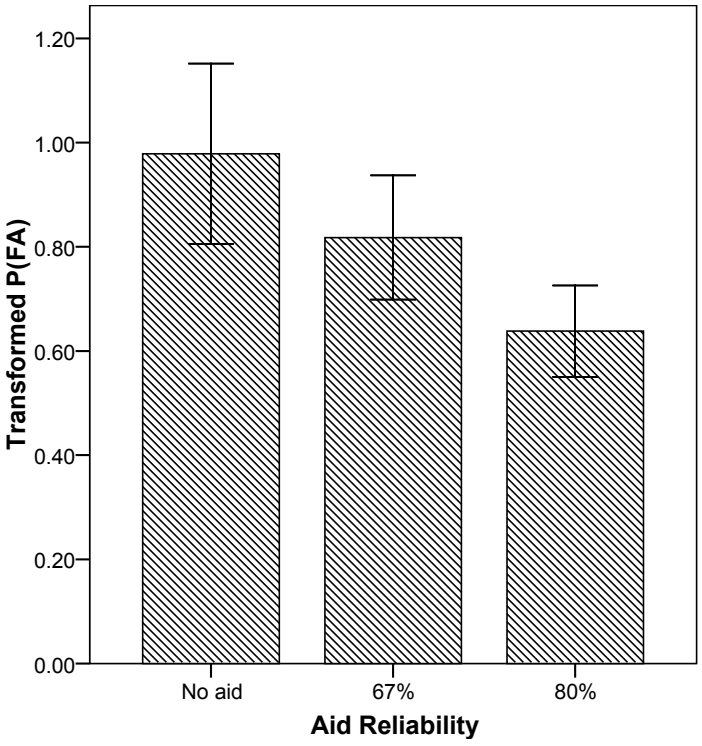


Figure 4. The effect of aid reliability on transformed P(FA)

[2] The assumption of sphericity was violated, the degrees of freedom were corrected using the conservative Greenhouse-Geisser value.

Miss (Miss Hostile Targets)

The 3 (aid reliability: no aid, 67%, 80%) X 2 (group: uninformed, informed) ANOVA on the transformed P(Miss) revealed no significant effects. The F values for each effect are: the effect of aid reliability, $F(2,44)=2.950$, $p=.063$; the effect of group, $F(1,22)=2.416$, $p=.134$, $r=.314$; the effect of the aid reliability X group interaction, $F(2, 44)=1.389$, $p=.260$.

Response Time

A 3 (aid reliability: no aid, 67%, 80%) X 2 (group: uninformed, informed) ANOVA were conducted using the response time (RT) as the dependent measure. There were no significant effects revealed in this analysis. The F values for each effect are: the effect of aid reliability, $F<1$; the effect of group, $F(1,22)=2.657$, $p=.117$, $r=.328$; the effect of aid reliability X group interaction, $F(2, 44)= 1.400$, $p=.257$.

Misuse and Disuse

Since the participants were informed that the 'friend' feedback (blue light) was always correct, they committed none or very few false alarm errors in the blue light trials. Therefore, the 'misuse and disuse' method was only used to analyze the participants' reliance on the 'unknown' feedback. For the two aided conditions, this section only reports the performance in those trials that the aid gave 'unknown' feedback (i.e. red light trials). For the no aid condition, this section reports the performance in all the trials, this serves as a baseline to compare with the false alarm rate (i.e. indication of misuse) and miss rate (i.e. indication of disuse) in the two aided conditions.

The 3 (aid reliability: no aid, 67%, 80%) X 2 (error type: FA, Miss) X 2 (group: uninformed, informed) ANOVA on the transformed P(Error) revealed a highly significant main effect of error type, $F(1, 22)=57.936$, $p<.001$, $r=.851$. In addition, a significant error type X group interaction was found, $F(1,22)=6.431$, $p=.019$, $r=.476$. As seen in Figure 5, both groups committed more false alarm mistakes than miss mistakes. However, because informed group made more false alarm errors but fewer miss errors than the uninformed group, the discrepancy between the two types of errors was larger in the informed group than the uninformed group. According to 'misuse and disuse' reliance measure method, this result indicated that the informed group relied on the 'unknown' feedback more than the uninformed group.

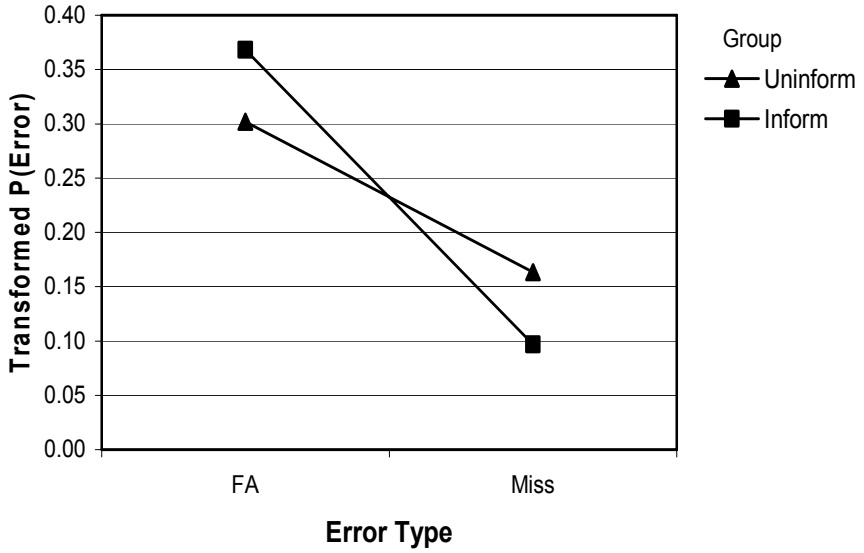


Figure 5. The error type X group interaction on transformed P(Error) in the red light trials

The aid reliability X error type interaction effect was found to be highly significant, $F(2,44)=9.521$, $p<.001$ (See Figure 6). Contrasts were performed comparing the no aid condition with the mean of the two aided conditions, and comparing the 67% reliability condition with the 80% reliability condition. There was a significant difference between the no aid condition and two aided conditions, $F(1,22)=10.773$, $p=.003$, $r=.573$, and a significant difference between the 67% reliability condition and the 80% reliability condition, $F(1,22)=6.558$, $p=.016$, $r=.479$. This interaction indicated that the effect of aid reliability had opposite effects for the different error types. As the aid reliability increased, the false alarm errors increased, but the miss errors decreased. This result suggests that the participants' reliance on the 'unknown' feedback increased with its reliability.

The other effects were not significant: the main effect of group, $F<1$; the main effect of aid reliability, $F(2, 44)=2.613$, $p=.085$; the reliability X group interaction, $F(2,44)<1$; the reliability X error type X group interaction, $F(2, 44)=1.175$, $p=.318$.

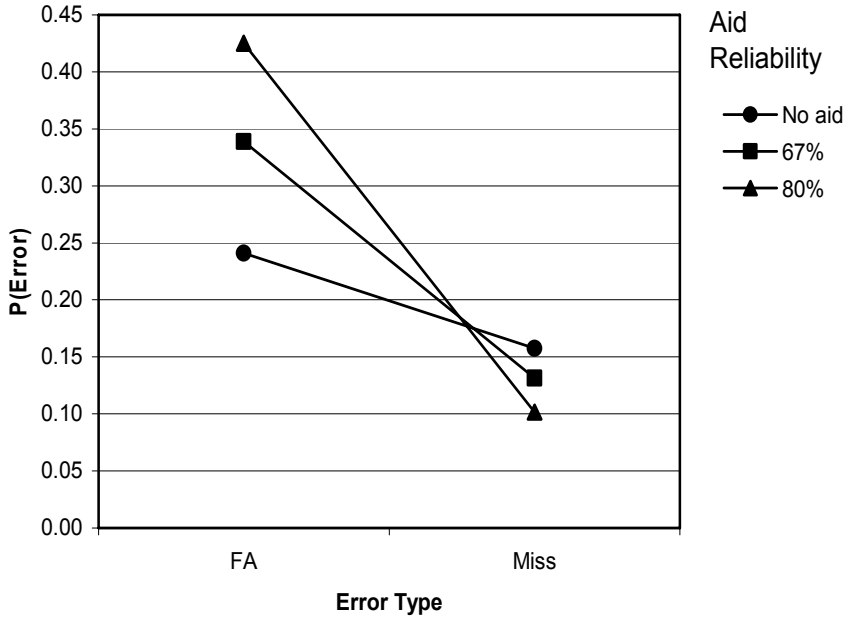


Figure 6. The aid reliability X error type interaction on transformed P(Error) in the red light trials

SDT Measures

Five participants did not make any miss errors in at least one of the three mission blocks. Because the SDT indices cannot be calculated if the miss error rate is zero, the data from those five participants was not included in the following SDT analysis.

Slope

The slopes of the empirically determined ROCs were calculated for the rest of the participants for each aid reliability condition (see Appendix M for the detailed calculation method). Since the participants always complied with the 100% correct 'friend' feedback, their performance in the blue light trials did not represent their own sensitivity well. Therefore, in the two aided conditions only the performance in the red light trials was used to calculate their ROCs. The one sample t-test revealed that there was no significant difference between the empirical slopes and

the null value 1.00, t(54) =1.295, p=.201 (2-tailed), Mean=1.196, SE=.151. This result indicated that d', C and β as general indices of the detection sensitivity and response bias were appropriate.

Sensitivity

The 3 (aid reliability: no aid, 67%, 80%) X 2 (group: uninformed or informed) ANOVA on sensitivity d' revealed no significant effects of aid reliability, F<1, group, F(1, 17)=1.150, p=.229, r=.252, or reliability X group interaction, F<1. Therefore, consistent with the hypothesis, participants' sensitivity didn't vary with the aid reliability or group assignment.

Decision Criterion C

The 3 (aid reliability: no aid, 67%, 80%) X 2 (group: uninformed, informed) ANOVA on the decision criterion revealed a significant main effect of aid reliability, F(2, 34)=5.272, p=.010 (see Figure 7). To break down this main effect, contrasts were performed comparing the no aid condition with the mean of the two aided conditions, and comparing the 67% reliability condition with the 80% reliability condition. There was a significant difference between the no aid condition and two aided conditions, F(1,17)=5.475, p=.032, r=.494, and a significant difference between the 67% reliability condition and the 80% reliability condition, F(1,17)=4.657, p=.046, r=.464. This result indicated that the participants' decision criterion was significantly lower when they received the 'unknown' feedback than when they did not receive that feedback. In addition, their decision criterion was lower when the 'unknown' feedback was 80% reliable than when it was 67% reliable, which suggested that they relied on the 'unknown' feedback more in the 80% reliability condition than the 67% reliability condition.

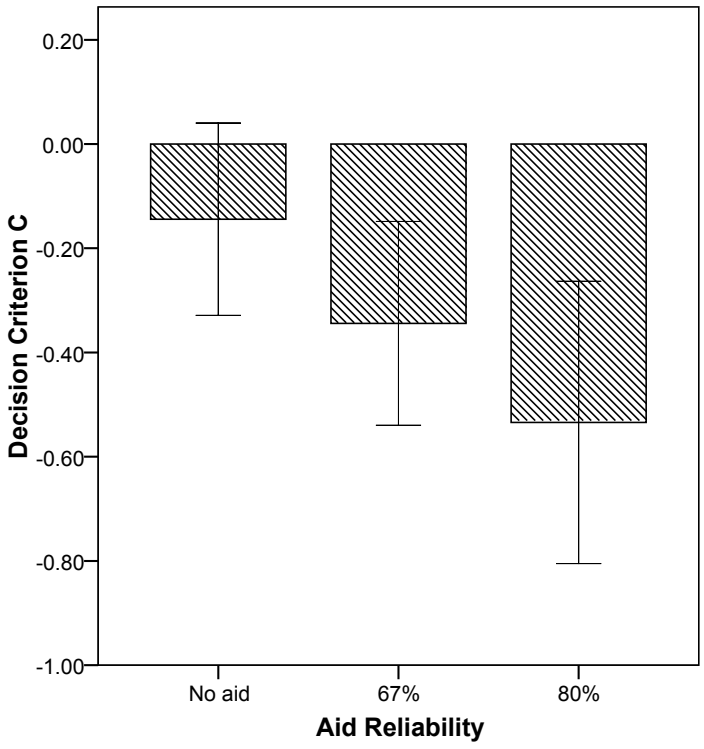


Figure 7. The main effect of aid reliability on decision criterion C

There was also a significant main effect of group, F(1, 17)=8.272, p=.010, r=.572. Regardless of aid reliability condition, the informed group (M=-.523) was more liberal in their decision to shoot than the uninformed group (M=.-138), which implied that the informed group relied on the

'unknown' feedback more than the uninformed group. No significant effect was found for the reliability X group interaction, F(2,34)=1.042, p=.364.

Appropriateness of Reliance

A Shapiro-Wilk test revealed that the assumption of normality was violated for bias ß. Therefore, a natural log transformation was applied to the bias ß. A series of one sample t-tests were conducted to test the difference between the ln ß and the optimal values, the results are listed in Table 3.

Table 3. T-test comparing participants' response bias with the optimal value

Group	Aid reliability	ln(Optimal Beta)	t-value	p-value (2-tailed)
Uninform	No aid	ln(1.00)	t(8)=.630	.546
Inform	No aid	ln(1.00)	t(9)=-2.568	.030
Uninform	67%	ln(0.50)	t(8)=3.538	.008
Inform	67%	ln(0.50)	t(9)=-.544	.600
Uninform	80%	ln(0.25)	t(8)=5.937	.000
Inform	80%	ln(0.25)	t(9)=-.063	.951

As seen in Figure 8, for the two aided conditions, the bias ß of participants in the uninformed group was significantly higher than the optimal value, while the bias ß of the informed group was not significantly different from the optimal value. Therefore, the informed group relied on the aid more appropriately than the uninformed group in the aided condition.

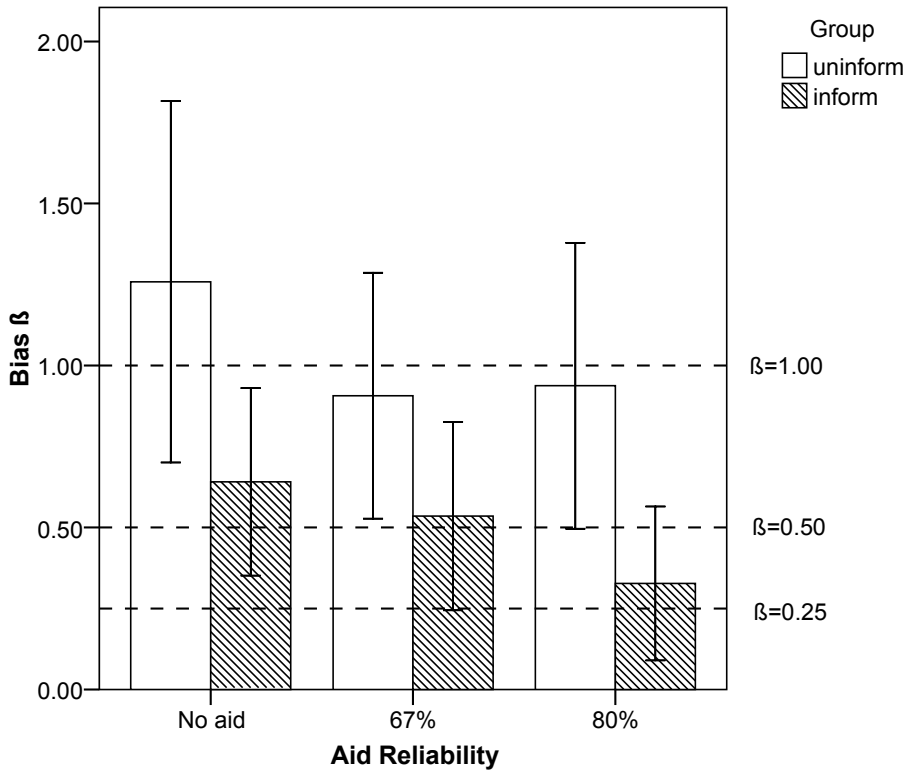


Figure 8. Mean and standard deviation of bias ß

For the no aid condition, the uninformed group did not deviate significantly from the optimal value, while the informed group was significantly lower than the optimal level. This result was unexpected. The stack histogram (see Figure 9) shows that two participants in the informed group were much more liberal than the others in the no aid condition. Their scores may have reduced the average ß of the informed group. When their data was ignored, the bias ß of the informed group in the no aid condition was not significantly different from the optimal value, t(7)=-1.877, p=.103.

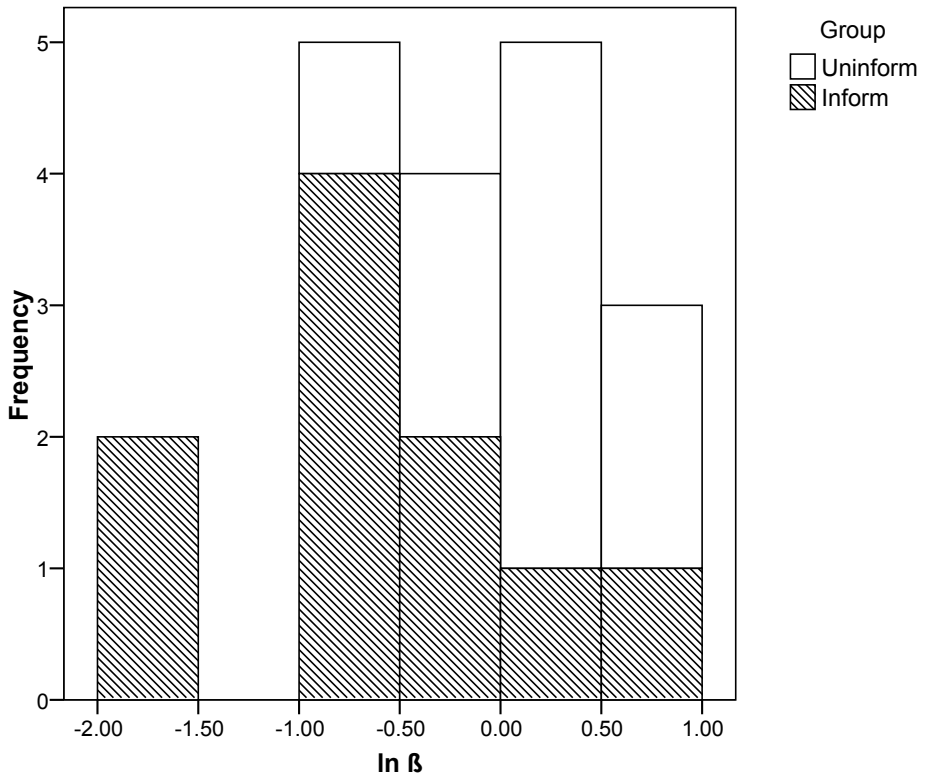


Figure 9. Stack histogram of the ln ß in the no aid condition

Subjective Rating

Trust

Trust in the Whole System

Trust in the whole system was rated on an empirically determined 7-point trust scale (Jian et al., 2000). The 2 (aid reliability: 67%, 80%) X 2 (group: uninformed, informed) ANOVA on the participants rating of trust in the whole system revealed a significant main effect of aid reliability, $F(1, 22)=7.183$, $p=.014$, $r=.496$. The participants trusted the whole system more in the 80% reliability condition ($M=3.74$) than 67% reliability condition ($M=4.30$). The main group effect and the aid reliability X group interaction effect were both not significant, $F<1$.

Trust in 'Friend' and 'Unknown' Feedback

Almost all the participants indicated absolute trust (rating 7) in the 'friend' feedback, as expected. However, their trust in the 'unknown' feedback (red light) varied (see Figure 10). A Shapiro-Wilk test revealed that the assumption of normality was violated for the trust ratings of the 'unknown' feedback. Several transformations were tested but none of them generated a normal

distribution. Therefore, non-parametric tests were used in analyzing the trust ratings on 'unknown' feedback. Figure 10 displays the mean and standard deviation of each condition.

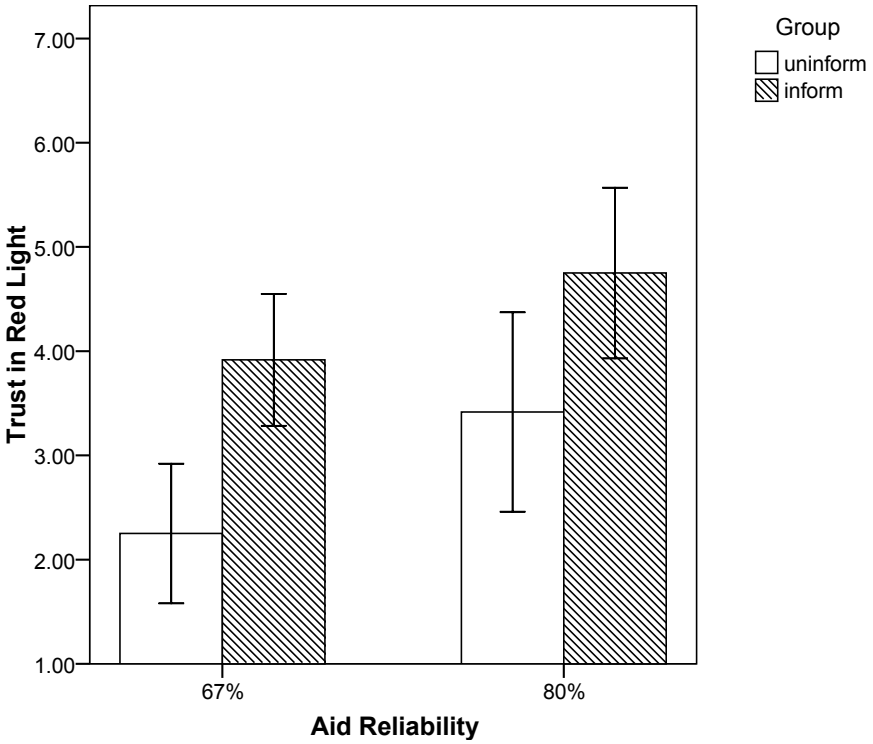


Figure 10. Mean and standard deviation of participants' trust in the 'unknown' feedback

The effect of aid reliability was significant in the informed group and the uninformed group. Both groups trusted the 'unknown' feedback significantly more in the 80% reliability condition than the 67% reliability condition, for the uninformed group, z=-2.232, p=.026, r=-.644, for the informed group, z=-1.997, p=.046, r =-.576.

The effect of group was significant in both aided conditions. The trust in the 'unknown' feedback was consistently higher in the informed group than the uninformed group, for the 67% reliability condition, U=19.50, p=.002, r=.633, for the 80% reliability condition, U=33.00, p=.018, r=-.482. In order to test whether the effect of group was larger in the 67% reliability condition than the 80% reliability condition, a Mann-Whitney test was performed on the difference of the trust ratings between the two aided conditions for each participant. No significant effect was revealed in this test, U=63.00, p=.590, r=-.110, which indicates that the effect group was similar in the two aided conditions.

Estimate of 'Unknown' Feedback Failure Rate

The participants' estimates of the 'unknown' feedback failure rate were compared with the real failure rate (see Figure 11). The one sample t-test comparing the participants' estimate in the 67% reliability condition with the real value 33% revealed no significant difference between these two values, t(11)=1.516, Mean=38.17%, p=.158 (2-tailed). Another one sample t-test comparing the participants' estimate in the 80% reliability condition with the real value 20% indicated that there was a significant difference between participants' estimate and real value, t(11)=2.721, Mean=31.25%, p=.020 (2-tailed).

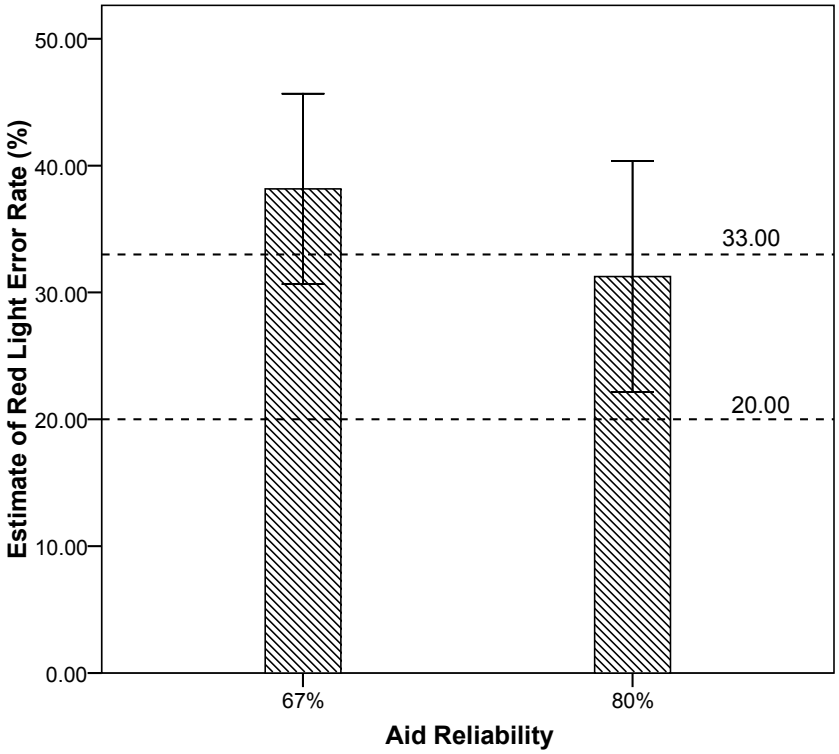


Figure 11. Mean and standard deviation of participants' estimation of red light failure rate

A paired samples t-test comparing the mean estimate of 'unknown' feedback failure rate between the two aided conditions indicated that there was no significant difference between the two conditions in the estimate value, $t(11)=1.634$, $p=.131$, $r=.442$, mean difference= 6.92%.

Overall, the uninformed group's estimate for the 67% reliable aid was not significantly different from the real value. However, they overestimated the failure rate for the 80% reliability aid. In addition, their estimates for the two reliability levels were not significantly different.

Relationships among Belief, Trust and Reliance

The estimates of 'unknown' feedback failure rate were collected from the participants who were not informed of this information. Therefore, the correlation analysis in this section was conducted only on the data from the uninformed group.

Trust & Estimation of 'Unknown' Feedback Failure Rate

Pooled over the two aided conditions (67% and 80%) for the uninformed group, a one-tailed Kendall's tau (τ) correlation analysis revealed that the participants' trust in the 'unknown' feedback was negatively correlated with their estimate of its failure rate, with a coefficient $\tau(20)=-.422$, $p=.011$. The coefficient of determination was $\tau^2=.178$, which indicates that participants' estimates of the 'unknown' feedback failure rate could account for 17.8 % of the variation in their trust in the 'unknown' feedback.

Reliance & Trust

Pooled over the two aided conditions (67% and 80%) for the uninformed group, a one-tailed Kendall's tau correlation analysis revealed that the participants' decision criterion C (an indication of their reliance on the 'unknown' feedback) was negatively correlated with the participants' trust in the 'unknown' feedback, with a coefficient $\tau(20)=-.570$, $p=.001$. The

coefficient of determination was τ^2=.325, which means that trust in the 'unknown' feedback could account for 32.5% of the variation in the participants' reliance on the 'unknown' feedback.

Reliance & Estimation of 'Unknown' Feedback Failure Rate

Pooled over the two aided conditions (67% and 80%) for the uninformed group, a one-tailed Kendall's tau correlation analysis revealed that the participants' decision criterion C was not significantly correlated with their estimate of the failure rate of the 'unknown' feedback, $\tau(20)$=.252, p=.072. The coefficient of determination was τ^2=.064, which means that estimation of the 'unknown' feedback failure rate may account for 6.4 % of the variation in the participants' decision criterion.

The above three correlation analyses also revealed that the coefficients of the first two correlations were much larger than the last one. It was likely that trust acted as an intermediate state in-between participants' belief about aid reliability and reliance action.

Discussion

The Relationship among Belief, Trust and Reliance

Previous research suggests that trust in an aid mediates the relationship between the operators' belief about the aid characteristics and their reliance on the aid (Lee & See, 2004). The results in this study support this claim. There is significant correlation between the participants' estimate of the 'unknown' feedback failure rate and their trust in the 'unknown' feedback, and between their trust in the 'unknown' feedback and reliance on the 'unknown' feedback. If the casual relationships between belief and trust, and between trust and reliance, do exist, the effects of the two independent variables in this experiment on the trust and reliance measures could be seen as a 'chain reaction' that started from their belief about the aid reliability.

The belief about the aid reliability was measured by the estimate of the 'unknown' feedback failure rate. Comparing the uninformed groups' estimates to the real value, their estimate for the 67% reliable aid was not significantly different from the real value. However, they underestimated the aid reliability for the 80% reliability aid. In addition, their estimates were not significantly different between the two reliability levels. This result partially supports the hypothesis that the uninformed group would not be able to correctly estimate the 'unknown' feedback failure rate based on their limited interaction with the aid. Because the participants had difficulties in estimating the failure rate, informing the participants of this information directly, as the informed group, was likely to benefit correct belief about the aid reliability.

For the trust in the 'unknown' feedback, both groups trusted the 'unknown' feedback more in the 80% reliability condition than the 67% reliability condition. In addition, the discrepancy between the two groups' belief about the 'unknown' feedback failure rate is reflected in the difference between the two groups' trust in the 'unknown' feedback. The uninformed group thought the 'unknown' feedback was less reliable than its real reliability, and they trusted the 'unknown' response less than the informed group who knew the real reliability.

As the trust, similar effects of aid reliability and reliability information were founded in the reliance on the 'unknown' feedback. Both groups relied on the 'unknown' feedback more often in the 80% reliability condition than the 67% reliability condition. However, regardless of the reliability conditions, the informed group relied on the 'unknown' feedback more than the uninformed group. Comparing with the optimal reliance level, in both aided conditions, the

informed group relied on the 'unknown' feedback appropriately, whereas the uninformed group did not rely on it often enough. It seems that the reliability information improved the appropriateness of reliance. In addition, since the reliance on 'unknown' feedback was significantly correlated with the trust in it, to some extent, it is reasonable to say that the reliability information also helped to engender appropriate trust.

Unlike the trust in 'unknown' feedback, the uninformed and informed groups had similar level of trust in the whole system, and they both had complete trust in the 'friend' feedback. The difference among the participants' trust in the whole system, the 'unknown' feedback, and the 'friend' feedback, reflects the functional specificity of trust (Lee & See, 2004). This might have been the result of the instruction about the distinctive reliability of the 'friend' and 'unknown' feedback. Also, the participants' trust in the 'unknown' feedback and the whole system was both higher in the 80% reliability condition than the 67% reliability condition. The participants discriminated in trusting the aids of different reliability, which reflects the resolution of their trust (Lee & See, 2004).

Comparison to Previous Studies

Identification Accuracy and Speed

In general, the previous studies did not find that the identification accuracy and speed was improved by the imperfect combat ID systems (Dzindolet et al., 2000, 2001a, 2001b; Karsh et al., 1995). Similarly, this experiment did not find significant differences in the speed of engagement decision and number of missing hostile targets among all the test conditions. However, in contrast to the previous studies, the combat ID aid in this study contributed to a significant reduction in the number of the friendly fire engagements. This improvement was found in both reliability levels, and it increased with the aid's reliability. In contrast to the suboptimal use behavior in the previous studies, the participants in this study generally relied on the aid reasonably. First, they almost always followed the 100% reliable 'friend' feedback. Second, they used the 'unknown' feedback to inform their identification decision but did not blindly follow it. Although the informed group relied on the 'unknown' feedback more appropriately than the uninformed group, this different reliance did not lead to a difference in the accuracy of their engagement decision. One possible reason might be the performance measure is less sensitive than the reliance measure, because it can be interfered by the potential individual difference in their identification sensitivity.

Effect of Aid Reliability Information on Reliance

Dzindolet et al. (2001b) examined whether providing the participants opportunities to use the combat ID aid or informing them of the aid reliability would be effective in generating appropriate reliance on the aid. They found that experience or instruction were not sufficient to make the participants rely on the feedback appropriately. Misuse of the aid was still prevalent. To some extent, the results in this experiment support that experience alone was not enough. The uninformed group was unable to correctly estimate difference between the two aid reliability levels, and overestimated the failure rate of the 80% reliable aid. Their inaccurate belief could lead to inappropriate trust and reliance on the aid. However, for the effect of explicitly informing the aid reliability, the results in this experiment are inconsistent with Dzindolet et al.'s (2001b) findings. In this experiment, while the uninformed group did not rely on the aid enough, the informed group relied on the aid appropriately.

The target proportion and aid reliability in Dzindolet et al.'s (2001b) study were the same as the 67% reliability condition in this study. Therefore, the different reliance should be resulted from other factors. The first possible factor is the participants' suspicion of the aid reliability instruction in Dzindolet et al.'s study (2001b). In their study, the 'ABSENT' feedback was always correct. However, even the participants who had been told this fact requested to view the slides again after receiving the 'ABSENT' feedback more than 20% of the time and did not follow the 'ABSENT' feedback about 9% of the time. The participants might be doubtful about the reliability information, and therefore the effect of this information was diminished. The second possible factor is the workload. In Dzindolet et al.'s study, other than detecting soldiers in the slides, the participants were also required to respond to audio stimuli. In addition, they were told that both tasks were equally important. Therefore, it is expected that the workload was higher in their study than in this experiment. Some research suggests that misuse of automation is more likely to happen when the participants are responsible for more tasks besides the automated task (Parasuraman et al., 1993).

Reliance Measure Methods

Both the 'misuse and disuse' method and the 'response bias' method were used to analyze the participants' reliance on the 'unknown' feedback in this experiment. Generally, the two different routes reached similar conclusions.

Comparing these two reliance analysis methods, there was two advantages of the 'response bias' method over the 'misuse and disuse' method. First, it allows the comparison between the participants' reliance level with the optimal level. Second, the 'response bias' method was easier to interpret than the 'misuse and disuse' method. In the 'response bias' method, the reliance was indicated by its measure of the response bias itself, whereas in the 'misuse and disuse' method, reliance was indicated by the contrast between the false alarm and miss errors. This means that, the main effects of aid reliability and group on reliance actually correspond to the error type X aid reliability interaction and error type X group interaction in the 'misuse and disuse' method. And the effect of aid reliability X group interaction on the reliance it would be shown in the three-way error type X aid reliability X group interaction. Higher order effects are usually hard to interpret, therefore, it is preferable to use a direct measure of reliance.

MILESTONE 3: EXPERIMENT II EXECUTION

This section describes the activities conducted during the Experiment II Execution phase of this project. The first part of this section describes the data collection. The second part gives a detailed overview of the results from Experiment II. The analysis methods were similar to Experiment I. The last part summarizes and discusses these results.

Data Collection

Participants

14 students with normal visual acuity from U of T were recruited in this experiment. Complete data were collected from 12 participants and only those data were used for analysis. Each participant was paid $30 CAD for their participation, and a bonus $10 CAD was given to the top performer who had the highest accuracy in engagement decisions.

Tasks and Procedures

The experiment took approximately two and half hours to complete. Each participant first went through a vision test, signed an informed consent form and filled out a demographic information survey. The participants were then given a sheet of instructions to explain the experimental procedure. Altogether there were four mission blocks each consisting of 60 trials. After each block the participants were asked to fill out two questionnaires (see Appendix H and Appendix I). The former was to measure their trust in the aid, and rating of the aid's usability. The later was to measure their workload. After they finished the four mission blocks, they were asked to complete another questionnaire about their preferences for the activation mode and feedback form (see Appendix J). **In contrast to Experiment I in which the participants only needed to shoot at the hostile targets, in this experiment they were required to kill a target if they considered it hostile.** That is, in Experiment I the participants could get a score even if they missed or just injured a terrorist, while in this experiment they had to kill a terrorist in order to get a score. This change was made to mimic the time pressure in the real-life situation. The participants had a better chance of killing a target if they could make a decision earlier.

Following the instructions, the experimenter showed the participants pictures of friendly and hostile targets. Then they went through a training session (120 trials). In the first half of the training session, the experimenter guided them to improve their identification skills; in the second half of the training session, the experimenter guided them to improve their shooting accuracy. After the training session, the participants started the four mission blocks. Before each block, they were informed of the activation mode and feedback form, as well as the aid reliability. The experimenter also asked them a list of questions to make sure that they understood the instructions correctly. For the complete instructions, please refer to Appendix G.

Measures and Instruments

The objective measures of the target identification performance, misuse and disuse, and SDT statistics were similar to Experiment I. In addition, because the participants could choose not to activate the aid or activate it more than once when the aid was in the manual mode, two supplementary measures were taken. They were:

- Activation rate: the percentage of trials that a participant pressed the activation button. This rate could reflect the disuse behavior.

$$P(\text{Activation}) = \frac{\text{Number of trials that a participant pressed the activation button}}{\text{Total number of trials}}$$

- Multi-Click rate: the percentage of trials that a participant activated the aid more than once. This rate indicated the inefficiency of the activation behavior.

$$P(\text{Multi-Click}) = \frac{\text{Number of trials that the aid was activated more than once}}{\text{Total number of trials}}$$

The calculation of the optimal bias ß was more complicated in this experiment than Experiment I. In this experiment, the participants were told that their score was the sum of the number of correct identification of friends and successful killing of terrorists. Therefore, unlike Experiment I, the value of correct identification of friend was not the same as the value of correct identification of terrorist, but the same as the value of killing of terrorist. In this experiment, when the participants decided to shoot, about 79.90% of the time they could successfully kill the targets. In other words, even if the participants correctly identified a terrorist, only 79.90% of the time they could get credit. Therefore, the value of correct identification of terrorist should be the value of killing of terrorist multiplied by the successful killing rate 79.90%.

Therefore, $\frac{V(CR)+C(FA)}{V(H)+C(M)} = \frac{1+0}{1*0.799+0} = 1.252$,

$$\beta_{optimal} = \frac{P(Friend \mid Unknown)}{P(Terrorist \mid Unknown)} * \frac{V(CR)+C(FA)}{V(H)+C(M)} = \frac{33\%}{67\%} * 1.25 = 0.626$$

For the subjective measures, the participants' trust in the aid, their impression of the aid's usability, and their workload were assessed. Trust was measured using the questions on 7 point scales as in Experiment I. Usability was evaluated from four perspectives – usefulness, easiness to use, easiness to learn, and satisfaction using 10 point scales (Lund, 2001). Workload was measured using the NASA Task Load Index (Hart & Staveland, 1988). In addition, the participants' preference for activation modes and feedback forms, as well as the reasons of their choice, were also recorded.

Data Analysis

Target Identification Performance

False Alarm (Friendly Fire)

The 2 (mode: auto, manual) X 2 (feedback: red light, no light) repeated-measures ANOVA was conducted on the transformed P(FA). The two main effects were not significant, $F<1$, and neither was the mode X feedback interaction, $F(1,11)=2.229$, $p=.164$, $r=.410$. Therefore, the false alarm rate was not affected by the activation mode or feedback form.

Miss (Miss Hostile Targets)

The 2 (mode: auto, manual) X 2 (feedback: red light, no light) repeated-measures ANOVA on the transformed P(Miss) revealed no significant main or interaction effects: the main effect of mode, $F(1,11)=1.066$, $p=.324$, $r=.297$; the main effect of feedback, $F<1$; the effect of the mode X feedback interaction, $F(1,11)=1.499$, $p=.246$, $r=.346$. Therefore, the miss rate was not affected by the activation mode or feedback form.

Response Time

The 2 (mode: auto, manual) X 2 (feedback: red light, no light) repeated-measures ANOVA was conducted using the response time (RT) as the dependent measure. The effect of mode was not significant, $F(1, 11)=3.907$, $p=.074$, $r=.502$, neither was the effect of feedback and the mode X feedback interaction were both non-significant, $F<1$.

Misuse and Disuse

Activation

A one-way repeated-measures ANOVA was conducted to examine the effect of feedback on the transformed P(Activation) when the aid was in the manual mode. There was no significant effect of feedback, $F<1$. Regardless of the 'unknown' feedback forms, the participants tended to activate the aid almost all of the time (M=92.9%).

The same one-way repeated-measures ANOVA using the transformed P(Multi-Click) as the dependent measure revealed a significant effect of feedback, $F(1,11)=5.294$, $p=.042$, $r=.570$. The participants activated the aid multiple times in one trial more frequently when the 'unknown' feedback (M=37.5%) was no light than when it was red light (M=17.3%).

Misuse and Disuse of 'Unknown' Feedback

The 2 (mode: auto, manual) X 2 (feedback: red light, no light) X 2 error type (FA, Miss) repeated-measures ANOVA on the transformed P(Error) in the 'unknown' feedback trials revealed no significant effects: the effect of error type, $F(1, 11)=4.033$, $p=.070$, $r=.518$; the mode X feedback interaction effect, $F(1, 11)=2.859$, $p=.119$, $r=.436$; the error type X mode interaction effect, $F(1, 11)=2.147$, $p=.171$, $r=.404$; the error type X feedback interaction effect, $F(1, 11)=1.884$, $p=.197$, $r=.382$; the rest of the effects, $F<1$.

Disuse of 'Friend' Feedback

Unlike Experiment I in which the participants almost always complied with the 'friend' feedback, in this experiment 10 out of 12 participants occasionally shot at targets in the 'friend' feedback trials.

The 2 (mode: auto, manual) X 2 (feedback: red light, no light) repeated-measures ANOVA on the transformed P(Error) of the 'friend' feedback trials revealed a significant effect of the feedback X mode interaction effect, $F(1,11)=7.606$, $p=.019$, $r=.639$. As seen in Figure 12, when the aid was in the auto mode, the influence of feedback form was minor. However, when the aid was in the manual mode, the disuse of 'friend' feedback was much severer in the red light feedback form than the no light form. The effect of feedback was not significant, $F(1,11)=4.098$, $p=.068$, $r=.521$, neither was the main effect of mode, $F<1$.

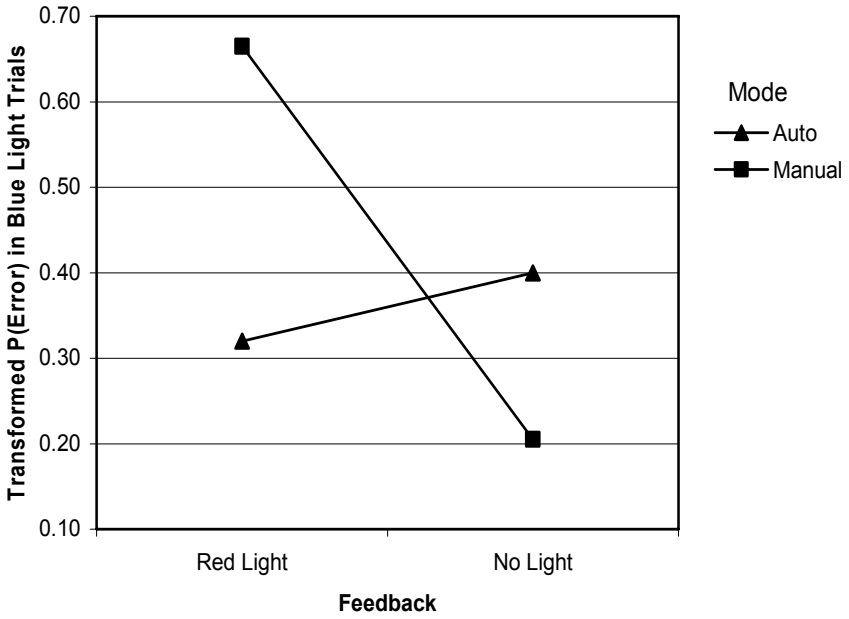


Figure 12. The mode X feedback Interaction on transformed P(Error) in the 'friend' feedback trials

SDT Statistics

Four participants did not make any false alarm or miss errors in at least one mission block. Because the SDT indices cannot be calculated if either the false alarm or miss error rate is zero, the data from these participants was not included in the following SDT analysis.

Slope

The slopes of the ROCs were calculated for the rest of the participants for each test condition (see Appendix M for a detailed illustration of the calculation method). As in Experiment I, only the participants' performance in the trials that the aid gave 'unknown' feedback was used to calculate their ROCs. A 2-tailed one sample t-test revealed that there was no significant difference between the empirical slope and the null value 1.00, $t(32) = 1.344$, $p = .189$, Mean=1.486, SE=.362. This result indicated that d', C and β as general indices of the detection sensitivity and response bias were appropriate.

Sensitivity

The 2 (mode: auto, manual) X 2 (feedback: red light, no light) repeated-measures ANOVA on the sensitivity d' revealed no significant effects: $F<1$ for the two main effects and the feedback X mode interaction.

Criterion

The 2 (mode: auto, manual) X 2 (feedback: red light, no light) repeated-measures ANOVA on the decision criterion C revealed no significant main or interaction effects: $F<1$ for all the effects. This result indicated that the participants did not rely on the 'unknown' feedback differently among all the test conditions.

Appropriateness of Reliance

A Shapiro-Wilk test revealed that the assumption of normality was violated for bias ß. Therefore, a natural log transformation was applied to the bias ß. A series of one sample t-tests were conducted to test the difference between the ln ß and the optimal values (see Table 4). As seen in

Figure 13, when the aid was in auto mode and no light feedback form, the participants' bias ß was significantly higher than the optimal value 0.626. This result indicates that the participants did not rely on the no light 'unknown' feedback enough when the aid was in auto mode.

Table 4. T-test comparing participants' response bias with the optimal value

Mode	Feedback	ln (Optimal Beta)	t-value	p-value (2-tailed)
Auto	Red Light	ln(0.626)	t(7)=1.307	.233
Auto	No light	ln(0.626)	t(7)=2.517	.040
Manual	Red Light	ln(0.626)	t(7)=.811	.444
Manual	No light	ln(0.626)	t(7)=-.038	.971

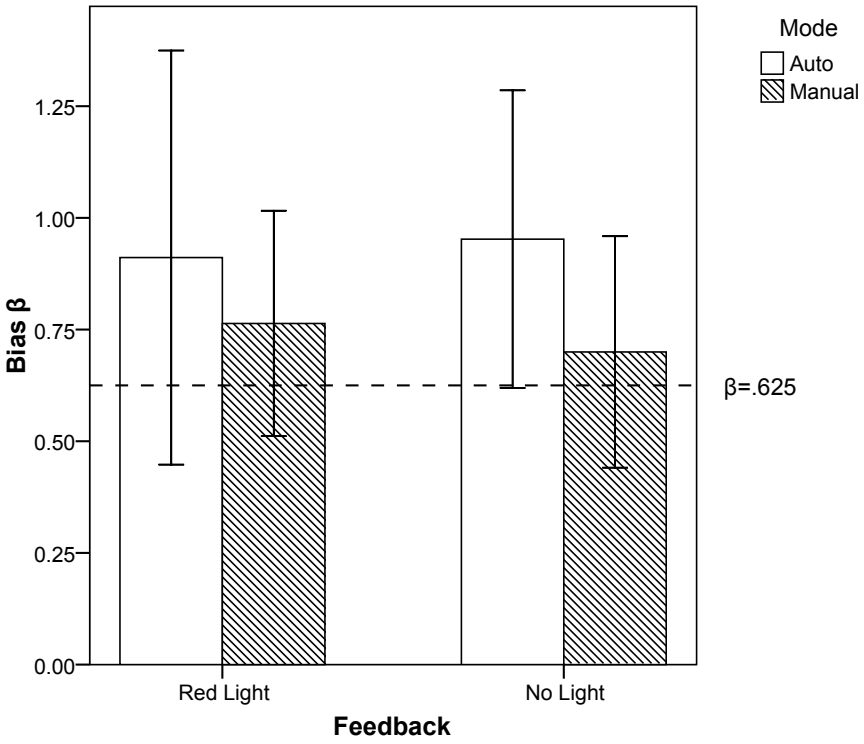


Figure 13. Mean and standard deviation of bias β

Subjective Rating

Trust

Trust in the Whole System

The 2 (mode: auto, manual) X 2 (feedback: red light, no light) repeated-measures ANOVA on the subjective ratings of trust in the whole system on revealed a highly significant main effect of feedback, $F(1, 11)=6.560$, $p=.026$, $r=.611$. This indicates that the participants trusted the whole system more when the 'unknown' feedback was no light ($M=6.36$) than when it was red light ($M=5.68$) The main effect of mode was not significant, $F<1$, neither was the mode X feedback interaction, $F(1, 11)=2.176$, $p=.168$, $r=.406$.

Trust in 'Friend' and 'Unknown' Feedback

Almost all the participants indicated absolute trust (rating 7) in the 'friend' feedback. However, their trust in the 'unknown' feedback varied. A Shapiro-Wilk test revealed that the assumption of

normality was violated for the trust ratings of the 'unknown' feedback. Several transformations were tested but none of them generated a normal distribution. So instead of the mixed design ANOVA, the non-parametric Wilcoxon Signed-Rank test was used to examine the effect of the activation mode in each feedback form condition, and to examine the effect of the feedback form in each activation mode condition.

The effect of activation mode was significant when the 'unknown' feedback form was no light. The participants trusted the 'unknown' feedback more in the automatic mode (M=2.42) than the manual mode (M2.08), z=-2.000, p=.046, r =-.577.

The other effects were not significant: for the effect of activation mode when the 'unknown' feedback form was red light, z=.000, p=1.000, r =.000; for the effect of feedback, in the auto mode, z=-.557, p=.564, r =-.167, in the manual mode, z=-1.667, p=.096, r =-.481.

Workload

The 2 (mode: auto, manual) X 2 (feedback: red light, no light) repeated-measures ANOVA on the subjective rating of workload revealed no significant effects: the effect of feedback, F(1,11)=4.386, p=.060, r=.534; the main effect of mode, F(1, 11)=3.079, p=.107, r=.468; the interaction effect, F<1.

Usability

The usability of the combat ID aid was assessed based on four aspects: usefulness, easiness to use, easiness to learn, and satisfaction. Figure 14 shows an overview of the result. Regardless of the activation mode and feedback form, the participants gave high ratings to the easiness to learn and easiness to use, and moderate ratings to usefulness and satisfaction. In addition, the auto mode with 'no light' feedback had the highest ratings in three of the four criteria; the manual mode with 'red light' feedback had the lowest ratings in three of the four criteria. Shapiro-Wilk tests revealed that the assumption of normality was violated for the usability ratings. Therefore, a series of Wilcoxon Signed-Rank tests were used to examine the effects of activation mode and feedback form for the four aspects of the usability ratings. No significant effects were found in the participants' ratings of these usability ratings.

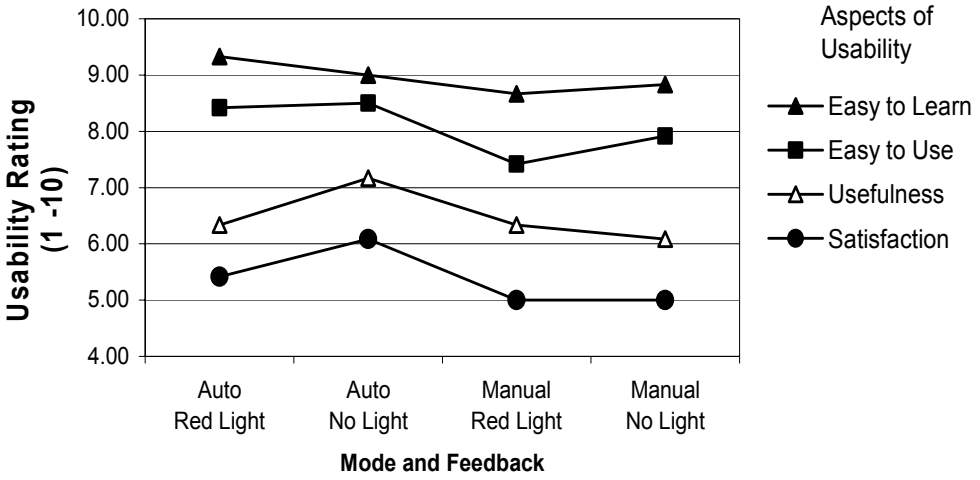


Figure 14. Subjective usability ratings on four aspects

Qualitative feedback

Nine out of the twelve participants preferred that the 'unknown' feedback was no light. Two primary reasons were given. First, the red lights hindered their focus on the targets. Second, even though the participants knew the red lights were fallible, most of them reported that upon seeing a red light, their first impulse was that the target was a terrorist and sometimes they just automatically shot at it. The other three participants liked the red light feedback and said that they were confused when the feedback did not include red lights. They could not be sure whether there was no light because the aid thought the target was a terrorist or because they did not successfully activate the automation.

Nine out of the twelve participants preferred the auto mode rather than the manual mode. They thought the auto mode was faster and easier to use than the manual mode. When the aid was in the auto mode, they did not need to worry about the activation button and could concentrate on shooting the targets. In contrast, two participants favored the manual mode because they felt they had more control of the recognition process. They could make a judgment by themselves first and then use the aid to confirm it. They thought it was better this way because the aid feedback could be misleading and distracting. One participant did not have a clear preference.

Relationship among Belief, Trust and Reliance

Trust and Belief

In this experiment, the participants' belief about the automation capabilities was reflected in their ratings in workload and usability. Pooled over the four conditions, the two-tailed Kendall's tau correlation analysis was conducted between trust ratings and workload ratings, and between trust ratings and usability ratings. The results are listed in Table 5. Both the trust in 'unknown' feedback and trust in the whole system was negatively correlated with workload ratings. This result suggests that the higher the workload, the less the participants trusted the 'unknown' feedback and the whole system. The coefficients of determination τ^2 were .090 and .101 respectively, which indicates that the workload ratings could account for 9.0 % of the variation in trust in 'unknown' feedback, and10.1% of the variation in trust in the whole system. In addition, trust in the whole system was significantly correlated with the satisfaction ratings. The coefficients of determination τ^2 were .152, which indicates that the satisfaction ratings could account for only 15.2 % of the variation in trust in the whole system.

Table 5. Kendall's tau correlation analysis between trust and belief (* p<.05)

		Workload	Usefulness	Easiness to Use	Satisfaction	Easiness to Learn
Trust in 'unknown' feedback	$\tau(32)$	-.300	-.200	-.261	-.036	-.195
	τ^2	.090	.040	.068	.001	.038
	Sig. (2-tailed)	.032*	.186	.086	.812	.186
Trust in whole system	$\tau(32)$	-.318	.184	-.224	.391	-.108
	T^2	.144	.034	.050	.153	.032
	Sig. (2-tailed)	.012*	.178	.104	.004*	.418

Reliance & Trust

Pooled over the four conditions, a one-tailed Kendall's tau correlation analysis revealed that the participants' trust in the 'unknown' feedback or the whole system was not significantly correlated with their reliance (i.e. decision criterion) on the 'unknown' feedback, τ (32)=-.117, p=.199, and τ (32)=-.090, p=.237. The coefficients of determination τ^2 were .014 and .008 respectively, which indicates that participants' trust in the 'unknown' feedback could only account for 1.4 % of the variation in their reliance, and their trust in the whole system could only account for 0.8% of the variation.

Reliance & Killing Rate

In this experiment, the appropriate reliance level depends not only on the reliability of the 'unknown' feedback, but also the success rate of killing a target. Therefore, the participants' reliance on the 'unknown' feedback might be affected by their success rate of killing a target. Pooled over the four conditions, the one-tailed Kendall's tau correlation analysis was conducted between the participants' reliance (i.e. decision criterion) and the participants' transformed success killing rate. No significant correlation was revealed in the analysis, τ (32) =.029, p=.410. The coefficients of determination τ^2 was .001, which suggest that the success rate of killing a target could only account for 0.1% variation in reliance.

Reliance & Belief

Pooled over the four conditions, the two-tailed Kendall's tau correlation analysis was conducted between the participants' reliance (i.e. decision criterion) on the 'unknown' feedback and workload, and between their reliance and usability ratings. The reliance was significantly correlated with satisfaction ratings, τ(32)=.283, p=.038. This result suggest that the more the participants satisfied with the aid, the more useful they thought the aid was, the more they relied on the aid. The satisfaction ratings could account for 8.0% of the variation in the reliance.

Discussion

Belief

In this experiment, all the participants were informed of the failure rate of the 'unknown' feedback. However, based on their subjective feedback, their belief about the aid characteristics was still different. When asked about their preferences of the aid's activation mode, most participants preferred to work with an aid in the auto mode because it was easier to activate and it allowed them to concentrate on visual identification. The participants also preferred an aid without the red light feedback to avoid interruption of their visual identification and biasing of their engagement decision. Despite the fact that the participants had preference for the activation mode and 'unknown' feedback form, no significant difference was found in their workload and usability ratings among the four conditions. These results might suggest that the workload and usability ratings are not sensitive to detect the participants' preference, or the participants' preference is not resulted from feelings in these rated aspects.

Trust

Although the participants were informed that the failure rate of the 'unknown' feedback was the same among the four conditions, they still trusted it differently. When the 'unknown' feedback was no light, their trust in the 'unknown' feedback was higher in the auto mode than the manual mode. This might be caused by the confusion when there was no feedback after manual activation. The participants reported that they were worried that they did not successfully

activate the aid. In fact, the participants tended to activate the aid more than once when the 'unknown' feedback was no light. In the auto mode, this confusion might be less because the activation was automatic. The ratings of trust show that the participants trusted the 'friend' feedback completely. In addition, they trusted the whole system more when the 'unknown' feedback was no light. This might be caused by the fact that most of the participants considered the red light feedback as disruptive and misleading, and they preferred an aid without red light feedback.

The results in trust ratings could be related to the notion that trust is not governed solely by the analytical process (Lee & See, 2004). In this experiment, the participants were aware that the reliability of the 'unknown' feedback was the same among all of the four conditions. If trust is completely based on the rational analysis, the participants should have trusted the 'unknown' feedback and the whole system equally in all of the four conditions. Therefore, other processes might influence the development of the participants' trust. The influence of the analogical and affective processes was to some extent supported by the significant relationships between the ratings of trust and the ratings of workload and usability, since these feelings about the aid might reflect the analogical and affective aspects of their trust. The higher the workload, the less the participants trusted the 'unknown' feedback and the whole system. The higher the satisfaction, the more the participants trusted the whole system. The low workload and high satisfaction might contribute to a positive feeling about the aid and lead to higher trust.

Reliance

Activation

When the aid was in the manual mode, no matter what the feedback form was, the participants activated the aid almost every time they spotted a target. This shows that they voluntarily tried to use feedback from the aid in their combat ID task. This result contradicts Karsh et al.'s (1995) finding that the participants only activated a combat ID aid occasionally. One cause of the different results might be that there was a delay of aid feedback in their experiment, whereas in this experiment, the aid feedback was immediately shown after activation. The delay of feedback increased the cost in using the aid and might have triggered the disuse behavior in Karsh et al.'s study. Another cause of the different activation behavior might be the potentially different self-confidence in conducting the experiment tasks in these two experiments. Since the manual accuracy was much better in their experiment than in this experiment, it is reasonable to expect that their participants had higher self-confidence. This high self-confidence might have led to the disuse of the aid in their experiment (Lee & Moray, 1994).

The participants' activation behavior was not very efficient – they often activated the aid more than once in a single trial. This inefficiency was severer when the aid did not send out a red light for hostile targets (P(multi_click) =.38) than when there was red light feedback (P(multi_click)=.17). The participants reported the confusion about receiving no feedback after manual activation. To make sure that the 'no light' response was not caused by unsuccessful activation, they sometimes re-aimed the weapon and activated the aid again.

Reliance on 'Unknown' Feedback

Both the 'misuse and disuse' and 'response bias' reliance measure methods indicate that the participants' reliance on the 'unknown' feedback was not significantly affected by activation mode and feedback form. The comparison between the participants' reliance on the 'unknown' feedback and the optimal reliance level shows that, when the aid was in the auto mode and the

'unknown' feedback was no light, the participants did not rely on the 'unknown' feedback often enough. This might be caused by the inconspicuousness of the 'unknown' feedback in this condition. The participants in this condition might sometimes not notice this implicit feedback. When the 'unknown' feedback was no light, it might be easier to notice it in manual mode than auto mode. Manual activation might remind the participants that the aid is sending feedback even though the feedback is implicit.

Unlike in Experiment I, there was no significant relationship between the trust in the 'unknown' feedback and the reliance on it in the current experiment. This result suggests that, in this experiment, there might be some other factors that have larger influence on reliance than trust, or the influence of trust on reliance might be overruled by some other factors. There was no significant relationship between the participants' reliance on 'unknown' feedback and their success rate of killing a target either. Because the instruction required the participants to kill the terrorist in order to get a score, ideally the participants should adjust their reliance according to their individual success rate of killing a target. This non-significant relationship suggests that the participants might not be able to estimate their own success rate of killing a target. Because this experiment did not collect the participants' estimate of their success rate of killing a target, it is not possible to test whether their estimate success rate of killing a target was significantly correlated with reliance.

The participants' reliance on the 'unknown' feedback was positively correlated with the ratings of satisfaction with the aid. The more satisfied the participants felt about the aid, the more they relied on the aid. Since the relationship between trust and reliance was not significant in this experiment, this result suggests the satisfaction feeling of the aid might have a direct influence on reliance, in other words, its effect on reliance might not be mediated by trust.

Reliance on 'Friend' Feedback

Unlike in Experiment I, the participants occasionally committed false alarm mistakes in the trials that the aid gave 'friend' feedback. This result was not anticipated because the participants had been informed that the 'friend' feedback was correct all the time and they also had absolute trust in it. The experimenter contacted the participants after the experiment to find out the causes for this unexpected result. The participants explained that they shot at the targets in the blue light trials not because they did not trust the 'friend' feedback but because they tried to react as quickly as possible, sometimes even before they saw the feedback from the aid. Therefore, the increased errors in blue light trials compared with Experiment I might be caused by the changed instruction about killing the hostile targets. The instruction might impose more time pressure on the participants than the instruction in Experiment I. The errors in blue light trials indicate that trust is not the only factor that affects the reliance behavior. Other factors, like time constraints, can intervene (Kirlik, 1993; Lee & See, 2004). The analysis of the false alarm rate in blue light trials also shows that, when the aid was in the manual mode, the participants made significantly more mistakes when the 'unknown' feedback was red light (Mean P(FA)=15.2%) than when it was no light (Mean P(FA)=3.4%). One possible explanation for this result is that, when the 'unknown' was no light, the participants might be more patient in waiting for feedback. As aforementioned, the no light 'unknown' feedback could cause confusion. The participants were cautious about it, and tended to activate the aid multiple times.

Summary

The purpose of this experiment was to find out whether improving the ecological validity of the simulated aid would result in changes in the participants' interaction with the aid. Based on the

findings in this experiment, the answer to this question was inconclusive. On the one hand, the participants' accuracy and speed in combat ID tasks and their reliance on the 'unknown' feedback was not affected by activation mode and feedback form. These results support the external validity of the conclusions from previous studies (Dzindolet et al., 2000, 2001a, 2001b; Karsh et al., 1995; Kogler, 2003) and Experiment I. One the other hand, the participants had a clear preference of these two interface features, and their trust in the 'unknown' feedback and the whole system were affected by these two interface features. These results suggest that the different interface features between the simulated aid and the real system prototypes have the potential to affect the reliance strategy, because many researches indicate that humans' reliance on the automation is affected by their belief about automation characteristics and trust in the automation (Lee & See, 2004; Lerch et al., 1997; Masalonis & Parasuraman, 2003; Muir, 1989).

The results in this experiment also illustrate that the effect of the trust in automation on the reliance behavior could be overruled or intervened by other factors. One power factor was the time constraints, even though the participants knew the 'friend' feedback was always correct and they trusted it completely, they sometimes did not succeed in relying on it because they tried to react fast. In addition, it seems that the participants' satisfaction with the aid could influence their reliance behavior directly without being mediated by trust.

REFERENCES

Bagheri, N., & Jamieson, G. A. (2004). The impact of context-related reliability on automation failure detection and scanning behaviour. *Systems, Man and Cybernetics, 2004 IEEE International Conference*

Bainbridge, L. (1983). Ironies of automation. *Automatica, 19*(6), 775-779.

Bourn, J. (2002). *Ministry of defence: Combat identification.* No. HC 661 Session 2001-2002. London, UK: National audit office. Retrieved Oct 10, 2006, from http://www.nao.org.uk/publications/nao_reports/01-02/0102661.pdf

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)**.** Melbourne, Australia.

Briggs, R. W., & Goldberg, J. H. (1995). Battlefield recognition of armored vehicles. *Human factors, 37*(3), 596-610.

Canadian killed by fellow soldier in Afghanistan shooting accident. (2006). *CBC News*, Retrieved March 6, 2007, from http://www.cbc.ca/world/story/2006/08/09/soldier-canadian.html

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)**,** Monterey, CA.

Corritore, C. L., Kracher, B., & Wiedenbeck, S. (2003). On-line trust: Concepts, evolving themes, a model. *International Journal of Human-Computer Studies, 58*(6), 737-758.

Defense science and technology plans (chapter VI) (2000). U.S. Department of Defense Deputy under Secretary of Defense.

Dixon, S. R., & Wickens, C. D. (2006). Automation reliability in unmanned aerial vehicle flight control: Evaluating a reliance-compliance model of automation dependence in high workload. *Human Factors. 48*(3), 474-486.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (1999). Misuse and disuse of automated aids. In *Proceedings of the 43rd Annual Meeting of the Human Factors and Ergonomics Society* (pp. 339-343). Santa Monica, CA: Human Factors and Ergonomics Society

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*(6), 697-718.

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.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., Dawe, L. A., & Anderson, B. W. (2001a). Predicting misuse and disuse of combat identification systems. *Military Psychology, 13*(3), 147-164.

Dzindolet, M. T., Pierce, L., Pomranky, R., Peterson, S., & Beck, H. (2001b). Automation reliance on a combat identification system. In *Proceedings of the 45th Annual Meeting of the Human Factors and Ergonomics Society* (pp. 532-536). Minneapolis/St. Paul, MN: Human Factors and Ergonomics Society

Field, A. P. (2005). *Discovering statistics using SPSS (2nd ed.)*. London: Sage.

Friendly fire: A recent timeline of fatal battlefield errors. (2007). *CBC News.* Retrieved March 2, 2007, from http://www.cbc.ca/news/interactives/tl-friendlyfire/

Frightening friendly fire facts. (2007). *Strategy Page*. Retrieved August 08, 2007, from http://www.strategypage.com/htmw/htmoral/articles/20070806.aspx

Frisconalti, M. (2005). *Friendly fire: The untold story of the U.S. bombing that killed four Canadian soldiers in Afghanistan.* Mississauga, Ontario: John Wiley & Sons Canada, Ltd.

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. Office of the Inspector General; United States: Department of Defense.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (task load index): Results of empirical and theoretical research. *Human Mental Workload, 1*, 139–183.

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

Hughes, B. K. (1996). *Combat identification*. Bellingham, Wash.: SPIE--the International Society for Optical Engineering.

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*(1), 53-71.

Jones, C. (1998). The battlefield combat identification system: A task force XXI response to the problem of direct fire fratricide. *ARMOR,* (January-February), 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, United States: Army Research Lab.

Kim, J. & Moon, J. Y. (1998). Designing emotional usability in customer interfaces - Trustworthiness of cyber-banking system interfaces. *Interacting With Computers*, *10*, 1-29.

Kirlik, A. (1993). Modeling strategic behavior in human-automation interaction: Why an "aid" can (and should) go unused. *Human factors, 35*, 221-242.

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. (2006). Human factors and ergonomics in automation design. In G. Salvendy (Eds.), *Handbook of human factors and ergonomics* (3rd ed.). New York: Wiley.

Lee, J. D., & Moray, N. (1994). Trust , self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies, 40*(1), 153-184.

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

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

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*. Cambridge, MA: MIT Press.

Lowe, C. (2007). *Cutting through the fog of war.* Retrieved Aug 20th, 2007, from http://www.defensetech.org/archives/003496.html

Lund, A. M. (2001). Measuring usability with the USE questionnaire. Usability Interface: Usability SIG Newsletter, October. Retrieved May 20th, 2007, from http://www.stcsig.org/usability/newsletter/0110_measuring_with_use.html

Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide*. New York, NY: 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 *Proceeding of the Human Factors and Ergonomics Society 48th Annual Meeting* (pp. 581-585). Santa Monica.CA: Human Factors and Ergonomics Society.

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.

Milne, G. R., Boza, M. E. (1999). Trust and concern in consumers' perceptions of marketing information management practices. *Journal of Interactive Marketing, 13* (1), 5–24.

Mosier, K. L., & Skitka, L. J. (1996). Human decision makers and automated decision aids: Made for each other? In R. Parasuraman, & M. Mouloua (Eds.), *Automation and human performance: Theory and applications*. Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.

Muir, B. M. (1989). *Operators' trust in and use of automatic controllers supervisory process control task.* Unpublished doctoral, University of Toronto, Ontario, Canada.

Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies. Special Issue: Cognitive engineering in dynamic worlds, 27*(5-6), 527-539.

Muir, B. M., & Moray, N. (1996). Trust in automation: 2. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics, 39*(3), 429-460.

Parasuraman, R., & Miller, C. A. (2004). Trust and etiquette in high-criticality automated systems. *Communications of the ACM, 47*(4), 51-55.

Parasuraman, R., Molloy, R., & Singh, I. (1993). Performance consequences of automation-induced "complacency". *International Journal of Aviation Psychology, 3*, 1-23.

Parasuraman, R., & Mouloua, M. (1996). In R. Parasuraman, M. Mouloua (Eds.), *Automation and human performance: Theory and applications*. Mahwah, NJ: Lawrence Erlbaum Associates.

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 & Cybernetics-Part A: Systems and Humans, 30*(3), 286-297.

Regan, G. (1995). *Back fire: The tragic story of friendly fire in warfare from ancient times to the gulf war*. London, UK: Robson Books.

Rierson, W. M., & Ahrens, D. A. (2006) Combat identification training: recognition of combat vehicles program. Countermeasure, Vol. 27, March 2006, published by the Army Combat Readiness Center at Fort Rucker, Alabama

Riley, V. (1994). *Human use of automation.* Unpublished doctoral, University of Minnesota, Minneapolis, MN.

Riley, V. (1996). Operator reliance on automation: Theory and data. In R. Parasuraman, & M. Mouloua (Eds.), *Automation and human performance: Theory and applications.* (pp. 19-35). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.

Roth, H. L., Lora, A. N. , & Heilman, K. M. (2002). Effects of monocular viewing and eye dominance on spatial attention. *Brain, 125*(9), 2023-2035.

Sheridan, T. (1996). Speculations on future relations between human and automation. In R. Parasuraman, & M. Mouloua (Eds.), *Automation and human performance* (pp. 449-460). Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Sheridan, T., & Parasuraman, R. (2006). Human-automation interaction. *Reviews of Human Factors and Ergonomics, 1*, 89-129.

Sheridan, T. B. (2002). *Humans and automation : System design and research issues*. [New York]: Wiley.

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). *Orlando, FL:* the International Society for Optical Engineering.

Sherman, K. B. (2002). Combat ID coming for individual soldiers. *Journal of Electronic Defense, 25(3)*, 34-35.

Shrader, C. R. (1982). *Amicicide: The problem of friendly fire in modern war.* (Research survey No.1). Fort Leavenworth, KS: Combat Studies Institute, U.S. Army.

Skitka, L. J., Mosier, K. L., & Burdick, M. (2000). Accountability and automation bias. *International Journal of Human-Computer Studies,* 52(4), 701-717.

Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does automation bias decision-making? *International Journal of Human-Computer Studies, 51*(5), 991-1006.

SIMLAS: Manworn combat training and ID system. (2006). RUAG Electronics. Retrieved October 23, 2006, from http://www.ruag.com/ruag/binary?media=99587&open=true

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

St. John, M., & Manes, D. I. (2000). Making unreliable automation useful. In *Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting* (pp. 332-336). Baltimore, MD: Human Factors and Ergonomics Society.

Steinweg, K. K. (1995). Dealing realistically with fratricide. *Parameters*, *Spring 1995*, 4-29.

US air strike kills Iraqi troops. (2007). *BBC News*, Retrieved March 6, 2007 http://news.bbc.co.uk/1/hi/world/middle_east/6346901.stm

U.S. Department of the Army. (1993). *Military Operations: U.S. Army Operations Concept for Combat Identification* (pp.1.). (TRADOC Pam 525-58). Fort Monroe, Va.: Training and Doctrine Command.

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

Wilson, K. A., Salas, E., & Priest, H. A. (2007). Errors in the heat of battle: Taking a closer look at shared cognition breakdowns through teamwork. *Human factors, 49*(2), 243-256.

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

Young, C. J. (2005). *Fratricide*. (Dispatches: Lessons learned for soldiers, 11 (1).) Kingston, ON, Canada: The Army Lessons Learned Center

APPENDICES

Appendix A: Experiment I – Informed Consent Form

INFORMED CONSENT FORM

Developing Human-Machine Interfaces to Support Appropriate Trust and Reliance on Automated Combat Identification Systems

Principal Investigator: M.A.Sc. Candidate Lu Wang
Faculty Supervisor: Professor Greg A. Jamieson
Department of Mechanical and Industrial Engineering
University of Toronto

This study is sponsored by Defense Research and Development Canada. The purpose of this study is to discover the necessary information that can facilitate soldiers' use of an automated Combat Identification (combat ID) system. The results of this study will guide the interface design for the combat ID system. This interface will help soldiers to better utilize the combat ID system, thereby reducing friendly fire incidents. You are invited to participate in this study because you are a student with normal or corrected-to-normal vision at the University of Toronto (U of T). The experiment will be conducted in the Cognitive Engineering Laboratory at U of T and there will be altogether 24 participants involved.

During the experiment, you will be seated in front of a computer workstation to interact with a combat ID virtual simulation and you will be asked to a) shoot hostile targets in simulated combat scene; and b) indicate your level of confidence in your judgment. The whole experiment will last approximately three hours, which includes the following sections:

1. Instruction (10 min): the investigator will give you instruction on how to complete tasks in the combat ID simulation.
2. Training (30 min): you will practice in training scenarios.
3. Formal Test (120 min): you will complete tasks in 3 mission blocks and answer short questionnaires.

The risk is minimal in this study and is comparable to playing a video computer simulated shooter game. You will receive a cash compensation of 7 CAD for every hour you spend on this study and an additional 9 CAD for completion of the whole experiment. In addition, you will have the potential to earn a bonus 10 CAD if you are the top performer among all the participants. The cash compensation will be paid to you right after the experiment. After we collect data from all the participants, you will be contacted and receive bonus if you are the one with the best performance.

Your privacy and identity will be carefully protected in this study. A Master List with your identity information will be kept in order to find and reward the participant with the best performance in the experiment. The Master List will be stored securely in a locked filing cabinet. Only the experimenters of this study and the financial officer in the MIE Department at U of T will have access to it. Once the experiment has been completed, the unidentifiable raw data of each participant will be assigned a "non-descriptive alias" and the Master List will be destroyed. In any publication, information will be provided in such a way that you cannot be identified.

Your participation in this study is completely voluntary. You may refuse to participate without any negative consequences. In addition, you may withdraw from the study at any time without any penalty, and request your data be destroyed. In that case your remuneration will be calculated based on the actual time you would have spent in the study, at a rate of 7 CAD per hour.

M.A.Sc. Candidate Lu Wang is undertaking this study in partial fulfillment of Master's Degree requirements. If you have any additional questions later about this study, Ms. Wang (lulu@mie.utoronto.ca, 416-978- 0881) will be happy to answer them. For information about participants' rights in scientific study, you can contact the Ethics Review Office at ethics.review@utoronto.ca or 416-946-3273.

You will be given a copy of this form to keep.

PARTICIPANT CERTIFICATION:

I have read this Informed Consent Form. I have had the opportunity to ask any questions that I had regarding the study, and I have received answers to those questions. By my signature I affirm that I agree to take part in this study as a research participant and that I have received a copy of this Informed Consent Form.

.....	.....
Signature of Research Participant	Signature of Investigator
.....	.....
(Please PRINT name)	(Please PRINT name)

.....
Date

Appendix B: Experiment I – Participant Information Survey

Vision:

Right: _____
Left: _____
Dominant eye: _____
Color blindness: _____

Age: _____

Sex: _____

Major: _____

How often do you play first-person shooter games?
A. Never B. Rarely C. Sometimes D. Regularly

Appendix C: Experiment I – Assessment of Instruction Comprehension

1. Please fill out the blank:

In each block, ___% of all targets will be Canadian soldiers.

When a target is a terrorist, the light on the combat ID aid should be _____.

When a target is a Canadian soldier, the light on the combat ID aid should be _____.

2. Please circle the right answer:

When a target is a terrorist, I should _____.
A. hold fire B. shoot it as soon as possible

When a target is a Canadian soldier, I should _____.
A. hold fire B. shoot it as soon as possible

The mistake of shooting a Canadian solider and the mistake of not shooting a terrorist are _____.
A. equally serious B. not equally serious

When the light on the combat ID aid is blue, it is ____ that the target is a terrorist.
A. possible B. not possible

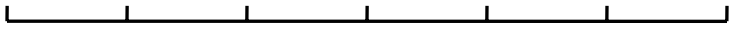
When the light on the combat ID aid is red, it is ____ that the target is a Canadian.
A. possible B. not possible

Appendix D: Experiment I – Trust and Reliability Estimation Questionnaire

Questionnaire after the Block with combat ID Aid

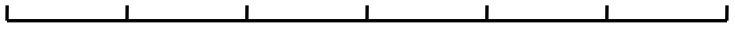
Please **circle the number** which best describes your feeling or your impression **in the mission block you just completed**. Remember, there are no right answers.

1. The aid is deceptive



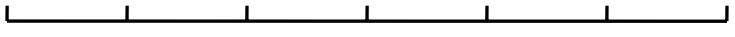
1 2 3 4 5 6 7
not at all extremely

2. The aid behaves in an underhanded (concealed) manner



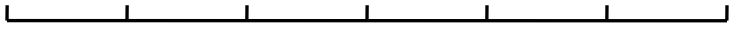
1 2 3 4 5 6 7
not at all extremely

3. I am suspicious of the aid's outputs



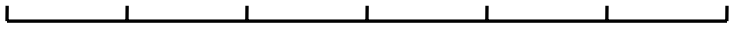
1 2 3 4 5 6 7
not at all extremely

4. I am wary of the aid



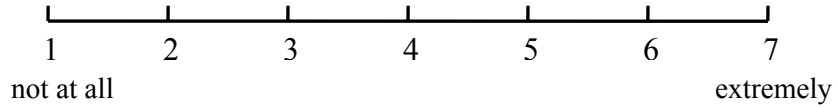
1 2 3 4 5 6 7
not at all extremely

5. The aid's action will have a harmful or injurious outcome

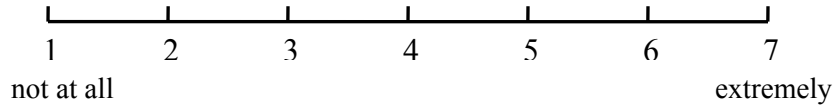


1 2 3 4 5 6 7
not at all extremely

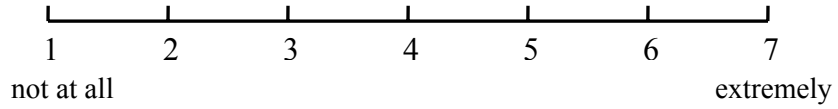
6. I am confident in the aid



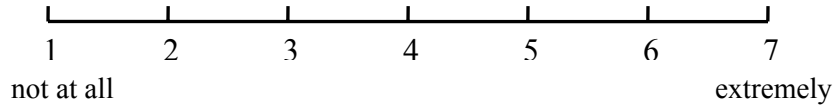
7. The aid provides security



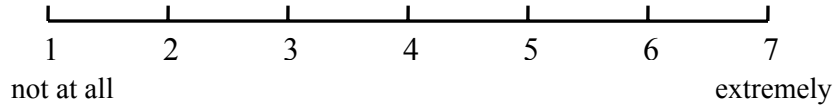
8. The aid is dependable



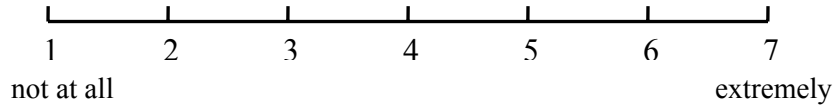
9. The aid is reliable



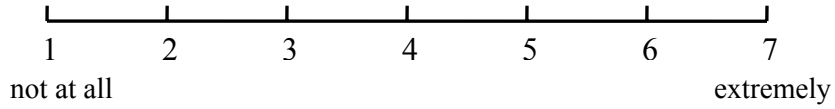
10. I can trust the aid



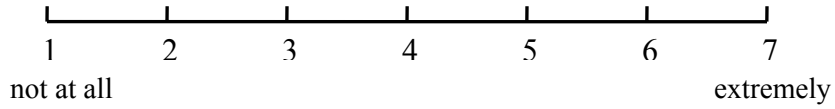
11. I am familiar with the aid



12. I can trust that **blue** lights indicate Canadian soldiers



13. I can trust that **red** lights indicate terrorists



14. I think _____% of the red lights were false (i.e. the targets were actually Canadian soldiers).

* Question 14 only appeared on the questionnaires for the uninformed group.

Appendix E: Experiment I – Instruction Scripts

Instruction 1: (experiment procedure)

You will complete 3 mission blocks in this experiment, each consisting of 120 trials. In each trial, an unknown soldier, which we call the "target", will appear in the simulated combat scene. These targets can be either hostile terrorists or friendly Canadian soldiers. Your task is to shoot (kill) terrorists **as soon as possible,** while holding fire on Canadian soldiers.

There are two types of error that can be made. One is made when you killed a friendly Canadian solider; the other is made when you didn't shoot a terrorist. Both errors are equally serious, and you should try to avoid them.

Your final score, which will determine whether you receive the bonus or not, will be calculated by the accuracy and speed of your response. After a target is killed or has run out of your sight, a screen will pop up to ask you to rate your confidence level in **your decision to shoot or hold fire.**

For the 120 trials in each block, the targets will be half terrorists and half Canadian soldiers. The order of the trials has been randomized. In 2 of the 3 mission blocks, you will have a combat identification (combat ID) aid to assist you. At the end of these 2 blocks, you will be asked to complete a short questionnaire.

Instruction 2: (combat identification system)

The combat ID aid in this experiment simulates a real-world combat ID system which comprises two parties, an interrogator and a transponder. As shown in the graph below, a soldier with an interrogator can send out an electronic message to another soldier, and if the second soldier is fitted with a compliant transponder he will send a message back to identify himself as a friend.

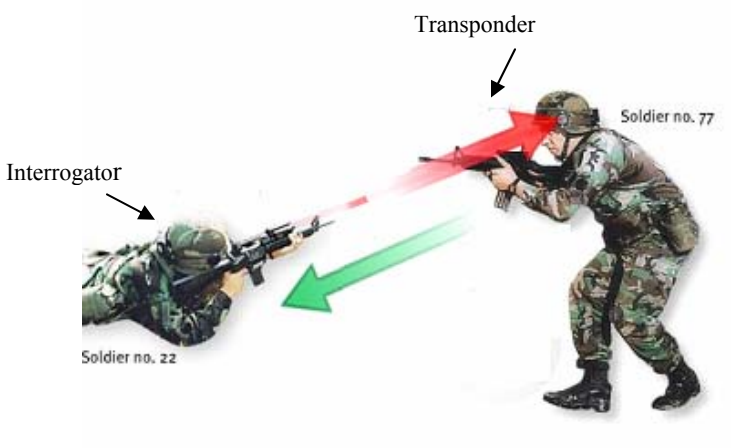


Figure 15. Interrogation process of the combat ID system

This interrogation process is simplified in the current simulation: you will not need to conduct the interrogation process; instead you will automatically receive feedback after your weapon is pointed at a target: a blue light indicates a Canadian soldier and a red light indicates a terrorist.

Figure 16. Feedback from the combat ID system

Although this aid is usually reliable, it is not 100% reliable all the time. This is because of the occasional failures in communications between an interrogator and a transponder in a chaotic battlefield. It is possible that a red light is shown when the target is actually friendly. In contrast, blue lights will always correctly identify Canadian soldiers: the blue light will never appear when the target is a terrorist.

In order to make sure you understand these instructions correctly, please answer the questions on the sheet of “Assessment of Instruction Comprehension”.

Instruction 3: (appearance of Canadian soldiers and terrorists)

The different appearance of terrorists and Canadian soldiers are illustrated in the graphs below. Note that they have different helmets, masks, weapons, etc. Please take your time to observe it and when you are ready we can move on to the training session.

Canadian **Terrorist**

Figure 17. Appearance of Canadian soldiers and terrorists

Instruction 4: (training)

The purpose of this training session is to develop your skill in identifying the targets and familiarize you with the simulation.

In the first 4 trials, the combat ID aid will be turned on. If you point your weapon to the targets, the light will indicate their identity. In these 4 trials, the response from the combat ID aid will be **always correct.**

The goal of the first 4 trials is to inform you with the targets' distinctive appearance. Therefore, Canadian soldiers and terrorists will follow a path sometimes very close to you, which will never happen in the mission blocks. Please **don't shoot** in these 4 trials, just carefully examine their different appearance.

After the first 4 trials, there will be 40 practice trials that are similar to the trials in the mission blocks. In these trials, the combat ID aid will be turned off. Therefore, you have to identify the target by yourself. Your task is to shoot (kill) terrorists as soon as possible, while holding fire on Canadian soldiers. A confidence scale will pop up at the end of a trial or after you kill a target. After your reply the confidence scale, the experimenter will tell you the target identity in the previous trial.

Instruction 5: (before each mission block)

Table 6. Instructions in different test conditions

Group	Reliability	Instrucition
Uninformed	**No aid**	Now you will start a mission block which will last about 35 minutes. In this mission block, the combat ID aid has been turned off, so you will not get any feedback from it.
Uninformed	**67%**	Now you will start a mission block which will last about 35 minutes. In this mission block, the combat ID aid has been turned on.
Uninformed	**80%**	Now you will start a mission block which will last about 35 minutes. In this mission block, the combat ID aid has been turned on.
Informed	**No aid**	Now you will start a mission block which will last about 35 minutes. In this mission block, the combat ID aid has been turned off, so you will not get any feedback from it.
Informed	**67%**	Now you will start a mission block which will last about 35 minutes. In this mission block, the combat ID aid has been turned on. **The possibility of the red lights being incorrect is about 33%.**
Informed	**80%**	Now you will start a mission block which will last about 35 minutes. In this mission block, the combat ID aid has been turned on. **The possibility of the red lights being incorrect is about 20%.**

Appendix F: Experiment II – Informed Consent Form

INFORMED CONSENT FORM

Developing Human-Machine Interfaces to Support Appropriate Trust and Reliance on Automated Combat Identification Systems

Principal Investigator: M.A.Sc. Candidate Lu Wang
Faculty Supervisor: Professor Greg A. Jamieson
Department of Mechanical and Industrial Engineering
University of Toronto

This study is sponsored by Defense Research and Development Canada. The purpose of this study is to discover the necessary information that can facilitate soldiers' use of an automated Combat Identification (Combat ID) system. The results of this study will guide the interface design for the Combat ID system. This interface will help soldiers to better utilize the Combat ID system, thereby reducing friendly fire incidents. You are invited to participate in this study because you are a student with normal or corrected-to-normal vision at the University of Toronto (U of T). The experiment will be conducted in the Cognitive Engineering Laboratory at U of T and there will be altogether 24 participants involved.

During the experiment, you will be seated in front of a computer workstation to interact with a Combat ID virtual simulation and you will be asked to a) shoot hostile targets in simulated combat scene; and b) indicate your level of confidence in your judgment. The whole experiment will last approximately three hours, which includes the following sections:

4. Instruction (10 min): the investigator will give you instruction on how to complete tasks in the CID simulation.
5. Training (30 min): you will practice in training scenarios.
6. Formal Test (120 min): you will complete tasks in 4 mission blocks and answer short questionnaires.

The risk is minimal in this study and is comparable to playing a video computer simulated shooter game. You will receive a cash compensation of 7 CAD for every hour you spend on this study and an additional 9 CAD for completion of the whole experiment. In addition, you will have the potential to earn a bonus 10 CAD if you are the top performer among all the participants. The cash compensation will be paid to you right after the experiment. After we collect data from all the participants, you will be contacted and receive bonus if you are the one with the best performance.

Your privacy and identity will be carefully protected in this study. A Master List with your identity information will be kept in order to find and reward the participant with the best performance in the experiment. The Master List will be stored securely in a locked filing cabinet. Only the experimenters of this study and the financial officer in the MIE Department at U of T will have access to it. Once the experiment has been completed, the unidentifiable raw data of each participant will be assigned a "non-descriptive alias" and the Master List will be destroyed. In any publication, information will be provided in such a way that you cannot be identified.

Your participation in this study is completely voluntary. You may refuse to participate without any negative consequences. In addition, you may withdraw from the study at any time without any penalty, and request your data be destroyed. In that case your remuneration will be calculated based on the actual time you would have spent in the study, at a rate of 7 CAD per hour.

M.A.Sc. Candidate Lu Wang is undertaking this study in partial fulfillment of Master's Degree requirements. If you have any additional questions later about this study, Ms. Wang (lulu@mie.utoronto.ca, 416-978- 0881) will be happy to answer them. For information about participants' rights in scientific study, you can contact the Ethics Review Office at ethics.review@utoronto.ca or 416-946-3273.

You will be given a copy of this form to keep.

PARTICIPANT CERTIFICATION:

I have read this Informed Consent Form. I have had the opportunity to ask any questions that I had regarding the study, and I have received answers to those questions. By my signature I affirm that I agree to take part in this study as a research participant and that I have received a copy of this Informed Consent Form.

.....	.....
Signature of Research Participant	Signature of Investigator
.....	.....
(Please PRINT name)	(Please PRINT name)

.....
Date

Appendix G: Experiment II – Instruction Scripts

Instruction 1: (experiment procedure)

You will complete 4 mission blocks in this experiment, each consisting of 60 trials. In each trial, an unknown soldier, which we call the "target", will appear in the simulated combat scene. These targets can be either hostile terrorists or friendly Canadian soldiers. Your task is to kill terrorists as soon as possible, while holding fire on Canadian soldiers.

There are two types of errors that can be made. One is made when you shoot at a friendly Canadian soldier; the other is made when you don't kill a terrorist. Both errors are equally serious, and you should try to avoid them.

Your final score, which will determine whether or not you receive the bonus, will be the total number of the trials that you hold fire on a Canadian soldier or successfully kill a terrorist. For each block, the targets will be half terrorists and half Canadian soldiers. The order of the trials has been randomized.

After a target is killed or has run out of your sight, a screen will pop up to ask you to rate your confidence in **your decision to shoot or hold fire.** After each block, you will be asked to complete a short questionnaire.

Instruction 2: (appearance of Canadian soldiers and terrorists)

Instruction 2 in Experiment II is identical to the Instruction 3 in Experiment I.

Instruction 3: (training)

The purpose of this training session is to develop your skills in identifying the targets, practice shooting and to familiarize you with the simulation.

First you will practice accurately shooting the target. You need to attempt to kill every target that appears on your screen. When the experimenter thinks that you've gained a certain level of accuracy you will move on to the second portion of training.

In the second portion of training, there will be 40 practice trials that are similar to the trials in the mission blocks. Your task is to kill terrorists as soon as possible, while holding fire on Canadian soldiers. A confidence scale will pop up at the end of a trial or after you kill a target. After you rate your confidence, the experimenter will tell you the correct target identity in the previous trial.

The tasks are hard, so please don't feel frustrated even if you make a lot of errors. The experimenter will help you to improve your performance.

Instruction 4: (combat identification system)

In the four mission blocks, you will have a combat identification (combat ID) aid to assist you. The combat ID aid simulates a real-world combat ID system which comprises two parties, an interrogator and a transponder. As shown in the picture below (Figure 15 was shown to the participants), a soldier with an interrogator can send out an electronic message to another soldier, and if the second soldier is fitted with a compliant transponder he will send a message back to identify himself as a friend.

Because of the occasional failures in communications between an interrogator and a transponder in a chaotic battlefield, it is possible that the system cannot recognize a Canadian as a friend. However, thanks to the encrypted code, it will never recognize a terrorist as a friend.

Instruction before each block:

Mode: Auto Feedback Form: Red Light

Now you will start a mission block which will last about 20 minutes. In this mission block, the combat ID aid is in the automatic mode. You will automatically receive feedback after your weapon is pointed at a target.

When the aid recognizes the target as a Canadian solider, it will show a blue light. Otherwise, a red light will be shown. Due to the communication errors mentioned earlier, it is possible that the aid displays a red light when the target is actually friendly. In contrast, the blue light will never appear when the target is a terrorist. During the interrogation, if the aid displays a red light, the possibility of it being incorrect is about 33%.

Figure 18. Feedback in auto mode & red light feedback form

Please orally answer the following questions:

When the combat ID aid responds a blue light, it is ____ that the target is a terrorist.

A. possible B. not possible

When the system responds a red light, it is ____ that the target is a Canadian soldier.

A. possible B. not possible

_____% of the "red light" responses will be false (i.e. the targets are actually Canadian soldiers).

Mode: Auto Feedback Form: No light

Now you will start a mission block which will last about 20 minutes. In this mission block, the combat ID aid is in the automatic mode. You will automatically receive feedback after your weapon is pointed at a target.

When the aid recognizes the target as a Canadian solider, it will show a blue light. Otherwise, **no light** will be shown. Due to the communication errors mentioned earlier, it is possible that the aid responds no light when the target is actually friendly. In contrast, the blue light will never appear when the target is a terrorist. During the interrogation, if the aid displays no light, the possibility of it being incorrect is about 33%.

Figure 19. Feedback in auto mode & no light feedback form

Please orally answer the following questions:

When the system responds no light, it is ____ that the target is a Canadian soldier.

A. possible B. not possible

When the combat ID aid responds a blue light, it is ____ that the target is a terrorist.

A. possible B. not possible

_____% of the "no light" responses will be false (i.e. the targets are actually Canadian soldiers).

Mode: Manual Feedback Form: Red Light

Now you will start a mission block which will last about 20 minutes. In this mission block, the combat ID aid is in the manual mode. If you want to interrogate a target, you need to point at the target and **press the "Alt" key**. The interrogation will last as long as the "Alt" key is depressed. If the Alt key is not pressed the aid will remain turned off. Please **only press and hold the "Alt" key each time when you want to interrogate a target**, in other words, don't depress it for the duration of the whole experiment.

When the aid recognizes a target as a Canadian solider, it will show a blue light; otherwise, a red light will appear. Due to the communication errors mentioned earlier, it is possible that a red light is shown when the target is actually friendly. In contrast, the blue light will never appear when the target is a terrorist. The possibility of the red lights being incorrect is about 33%.

Figure 20. Feedback in manual mode & red light feedback form

Please orally answer the following questions:

When the combat ID aid responds a blue light, it is ____ that the target is a terrorist.

A. possible B. not possible

When the system responds a red light, it is ____ that the target is a Canadian soldier.

A. possible B. not possible

_____% of the "red light" responses will be false (i.e. the targets are actually Canadian soldiers).

Mode: Manual Feedback Form: No light

Now you will start a mission block which will last about 20 minutes. In this mission block, the combat ID aid is in the manual mode. If you want to interrogate a target, you need to point at the target and **press the "Alt" key**. The interrogation will last as long as the "Alt" key is depressed. If the Alt key is not pressed the aid will remain turned off. Please **only press and hold the "Alt" key each time when you want to interrogate a target**, in other words, don't depress it for the duration of the whole experiment.

When the aid recognizes a target as a Canadian solider, it will show a blue light; otherwise, **no light** will be shown. Due to the communication errors mentioned earlier, it is possible that the aid responds no light when the target is actually friendly. In contrast, the blue light will never appear when the target is a terrorist. During the interrogation, if the aid displays no light, the possibility of it being incorrect is about 33%.

Figure 21. Feedback in manual mode & no light feedback form

Please orally answer the following questions:

When the system responds no light, it is ____ that the target is a Canadian soldier.

A. possible B. not possible

When the combat ID aid responds a blue light, it is ____ that the target is a terrorist.

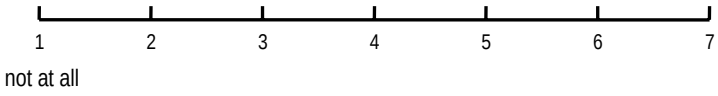
A. possible B. not possible

_____% of the “no light” responses will be false (i.e. the targets are actually Canadian soldiers).

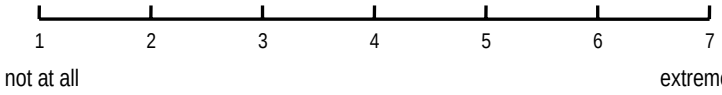
Appendix H: Experiment II – Trust and Usability Questionnaire

Please **circle the number** which best describes your feeling or your impression **in the mission block you just completed**. Remember, there are no right answers.

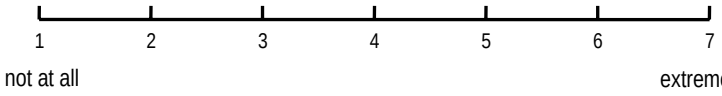
1. The aid is deceptive



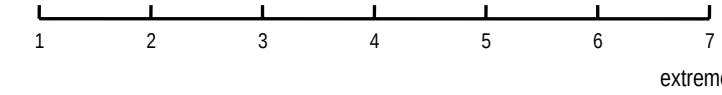
2. The aid behaves in an underhanded (concealed) manner



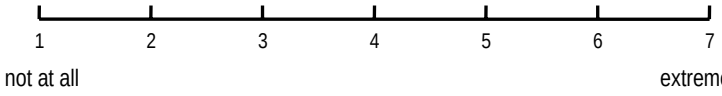
3. I am suspicious of the aid's outputs



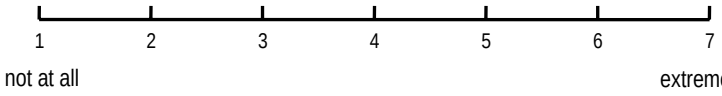
4. I am wary of the aid



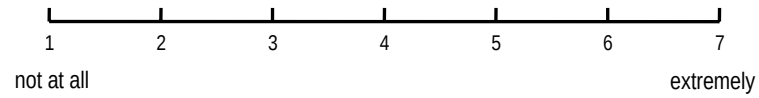
5. The aid's action will have a harmful or injurious outcome



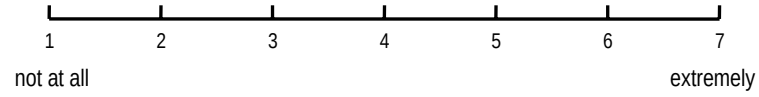
6. I am confident in the aid



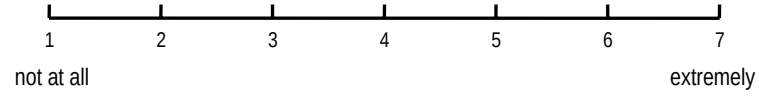
7. The aid provides security



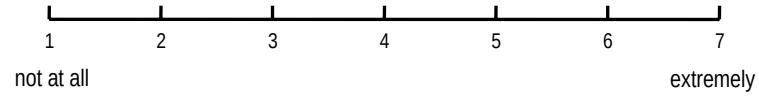
8. The aid is dependable



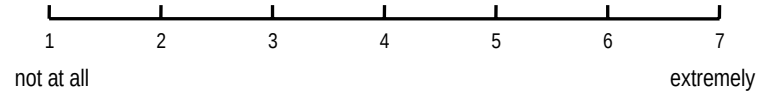
9. The aid is reliable



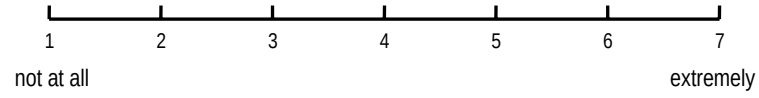
10. I can trust the aid



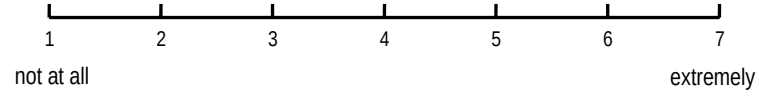
11. I am familiar with the aid



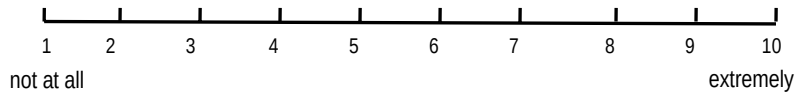
12. I can trust that blue lights indicate Canadian soldiers



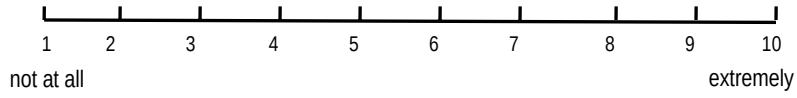
13. I can trust that “no light” / “red light” indicate terrorists



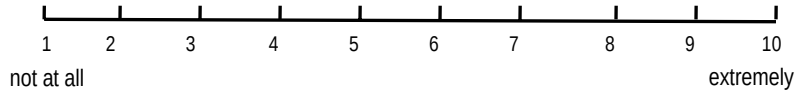
14. The aid helps me be more accurate in target identification.



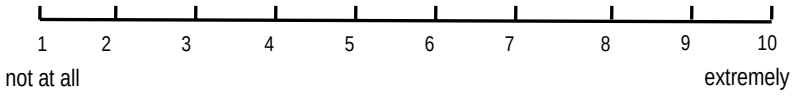
15. The aid is easy to use.



16. I easily remember how to use the aid.



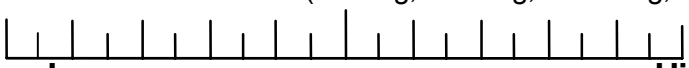
17. I am satisfied with the aid.



Appendix I: Experiment II – NASA TLX Questionnaire

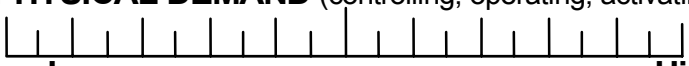
Rate the trial by marking each scale at the point which matches your experience. Each line has two endpoint descriptors to help describe the scale. Please consider your responses to these scales carefully.

MENTAL DEMAND (thinking, deciding, searching, remembering)



Low (easy, simple) **High** (demanding, complex)

PHYSICAL DEMAND (controlling, operating, activating)

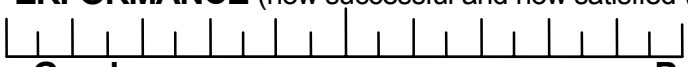


Low (easy, restful) **High** (demanding, laborious)

TEMPORAL DEMAND (time pressure)

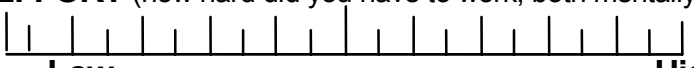
Low (leisurely) **High** (frantic)

PERFORMANCE (how successful and how satisfied were you with performing this task?)



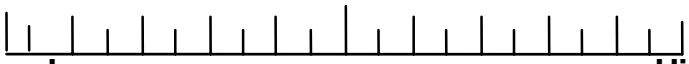
Good **Poor**

EFFORT (how hard did you have to work, both mentally and physically?)



Low **High**

FRUSTRATION



Low (gratified, complacent) **High** (discouraged, annoyed)

NASA-TLX WEIGHTINGS

WHICH FACTOR WAS THE MOST IMPORTANT TO YOU DURING THE TASK (A OR B)?

Definitions

Mental Demand (thinking, deciding, searching, remembering)
Physical Demand (controlling, operating, activating)
Temporal Demand (time pressure)
Performance (how successful and how satisfied were you with performing this task)
Effort (how hard did you have to work, both mentally and physically)
Frustration (level of frustration while performing this task)

Circle the factor that is more important to you.

A	B
Mental Demand	Physical Demand
Mental Demand	Temporal Demand
Mental Demand	Performance
Mental Demand	Effort
Mental Demand	Frustration
Physical Demand	Temporal Demand
Physical Demand	Performance
Physical Demand	Effort
Physical Demand	Frustration
Temporal Demand	Performance
Temporal Demand	Effort
Temporal Demand	Frustration
Performance	Effort
Performance	Frustration
Effort	Frustration

Appendix J: Experiment II – Preference Questionnaire

Please **fill out the blank** based on your feeling or your impression **in all the mission blocks**. Remember, there are no right answers.

I prefer to use the aid in _____ mode to _____ mode,
A. automatic B. manual

Reason:
_____
_____
_____
_____

I prefer to use the aid _____ to _____.
A. with red light indication B. without red light indication

Reason:
_____
_____
_____
_____

Appendix K: Ethics Review Approval

UNIVERSITY OF TORONTO
Office of the Vice-President, Research and Associate Provost

Ethics Review Office

PROTOCOL REFERENCE #19278 January 23, 2007

Prof. G. A. Jamieson
Mechanical & Industrial Engineering
5 King's College Road
University of Toronto
Toronto, ON M5S 3G8

Ms. L. Wang
Mechanical & Industrial Engineering
5 King's College Road
University of Toronto
Toronto, ON M5S 3G8

Dear Prof. Jamieson and Ms. Wang:

Re: Your research protocol entitled, "Developing Human-Machine Interfaces to Support Appropriate Trust and Reliance on Automated Combat Identification Systems" (Revised version received Jan. 22, 2007) by Prof. G. A. Jamieson (supervisor), Ms. L. Wang (Master's student)

ETHICS APPROVAL	**Original Approval Date: January 23, 2007** **Expiry Date: January 22, 2008**

We are writing to advise you that a member of the Social Sciences and Humanities Research Ethics Board has granted approval to the above-named research study, for a period of **one year**, under the REB's expedited review process. Ongoing projects must be renewed prior to the expiry date.

This approval has been issued with the understanding that all other appropriate approvals (where applicable) have been sought. Copies of valid approval letters from other relevant institutions should be submitted as soon as possible.

The following documents (revised versions received Jan. 22, 2007) have been approved for use in this study: Consent Form, Instruction Script, Questionnaire and Recruitment Flyer (to be printed on U of T departmental letterhead). Participants should receive a copy of their consent form.

Any changes to the approved protocol or consent material must be reviewed and approved through the amendment process prior to its implementation. Any adverse or unanticipated events should be reported to the Ethics Review Office as soon as possible.

Best wishes for the successful completion of your project.

Yours sincerely,

Marianna Richardson

Marianna Richardson
Ethics Review Coordinator

Simcoe Hall, 27 King's College Circle, Room 10A, Toronto Ontario M5S 1A1
TEL: 416/ 946-3273 FAX: 416/ 946-5763 EMAIL: ethics.review@utoronto.ca

Appendix L: Experiment I – Participant Vision and Demographic Data

Table 7. Vision and demographic data in Experiment I

No.	Dominant Eye	Right Eye Vision	Left Eye Vision	Age	Sex	Major	Frequency of playing shooter game
1	R	10/7.5	10/7.5	25.00	M	Mechanical Engineering	Sometimes
2	R	10/10	10/10	28.00	M	Mechanical Engineering	Rarely
4	R	10/10	10/10	32.00	M	Pharmacology	Rarely
5	R	10/7.5	10/10	33.00	F	Medicine	Never
6	R	10/7.5	10/7.5	25.00	F	MIE	Never
7	R	10/7.5	10/7.5	19.00	M	material science engineering	regularly
8	R	10/10	10/10	28.00	M	Human Factors	Rarely
9	R	10/7.5	10/10	34.00	M	Industrial Engineering	Sometimes
10	R	10/7.5	10/7.5	27.00	M	Industrial Engineering	Never
11	L	10/7.5	10/7.5	26.00	M	Engineering	Sometimes
12	R	10/7.5	10/7.5	19.00	M	Human Biology	regularly
13	L	7.5/10	10/10	19.00	M	Mechanical Engineering	Sometimes
15	R	10/10	10/12.5	24.00	M	Computer	Sometimes
16	R	10/7.5	10/7.5	20.00	F	Human Biology / Psychology	Rarely

17	R	10/7.5	10/7.5	24.00	M	Computer Science	Sometimes
18	R	10/7.5	10/7.5	33.00	M	Forestry	Sometimes
19	R	10/10	10/7.5	19.00	M	Mechanical Engineering	regularly
20	R	10/7.5	10/10	19.00	F	Physiology	Rarely
21	R	10/10	10/10	23.00	F	math	Never
22	L	10/7.5	10/7.5	18.00	M	Computer Science	Sometimes
23	L	10/10	10/10	22.00	M	Urban Studies	Sometimes
24	R	10/7.5	10/7.5	20.00	M	Mechanical Engineering	Rarely

Appendix M: Experiment I – Calculation of the ROC Slope

To obtain the ROC slope for each participant under each aid reliability condition, we generally followed the calculation method specified in a previous similar study (Dzindolet et al., 2001a).

Step 1: Cumulative Response Matrix

An overall cumulative response matrix was determined for each participant beginning with a highly confident response that a target was a Canadian and proceeding through the opposite extreme of a highly confident response that a target was a terrorist.

Table 8. A sample participant's (#12) cumulative response matrix for the no aid conditions

	Target	Canadian					Terrorist					
	Confidence Rating	5	4	3	2	1	1	2	3	4	5	Total
Response frequency	Canadian	4	29	2	0	0	1	2	5	15	2	60
	Terrorist	2	12	5	0	0	0	2	11	5	23	60
Response probability	Canadian	0.07	0.48	0.03	0.00	0.00	0.02	0.03	0.08	0.25	0.03	0.50
	Terrorist	0.03	0.20	0.08	0.00	0.00	0.00	0.03	0.18	0.08	0.38	0.50
Cumulative probability	Canadian	1.00	0.93	0.45	0.42	0.42	0.42	0.40	0.37	0.28	0.03	
	Terrorist	1.00	0.97	0.77	0.68	0.68	0.68	0.68	0.65	0.47	0.38	
Cumulative Z-score	Canadian		1.50	-0.13	-0.21	-0.21	-0.21	-0.25	-0.34	-0.57	-1.83	
	Terrorist		1.83	0.73	0.48	0.48	0.48	0.48	0.39	-0.08	-0.30	

Note: Confidence Rating: 1 – not at all confident; 2 – slightly confident; 3 – somewhat confident; 4 – confident; 5 – highly confident

Step 2: Plot Empirically Determined ROC

We then transformed and plotted these cumulative proportions onto the z-axis, which represented empirically determined ROCs. The slope of the ROC plotted in standard coordinates was determined through the method of least squares, which could resulted in the least amount of difference between the observed data and the regression line.

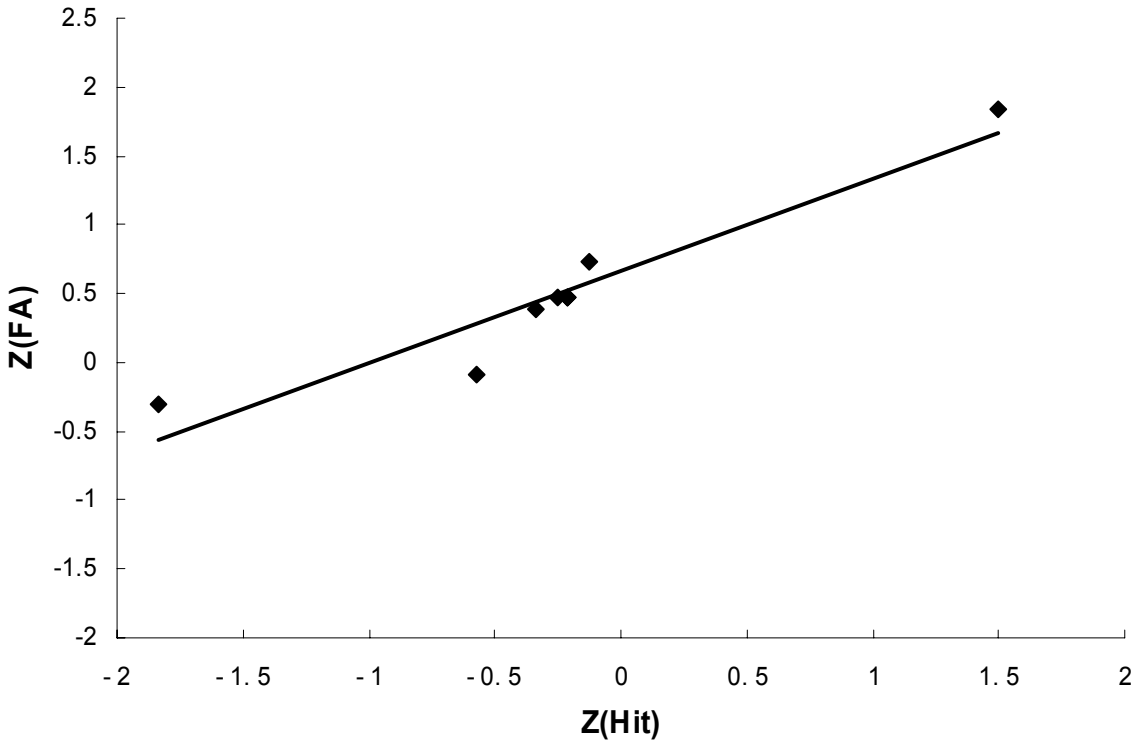


Figure 22. Sample ROCs for Participant #7 for the no aid condition

Appendix N: Experiment II – Participant Vision and Demographic Data

Table 9. Vision and demographic data in Experiment II

No.	Dominant Eye	Right Eye Vision	Left Eye Vision	Age	Sex	Major	Frequency of playing shooter game
1	L	10/10	10/10	23.00	F	Industrial Engineering	Never
2	R	10/10	10/10	28.00	M	Computer Science	Sometimes
4	R	10/7.5	10/10	22.00	F	Zoology	Never
5	R	10/7.5	10/10	21.00	M	Bioethics/ Religion	Rarely
6	R	10/12.5	10/15	20.00	F	Math	Never
7	R	10/12.5	10/10	23.00	M	Computer Science	Sometimes
8	R	10/12.5	10/7.5	21.00	F	Engineering Science	Never
9	R	10/12.5	10/10	38.00	M	Psychology/ Sociology	Rarely
10	R	10/10	10/7.5	22.00	F	Sociology	Sometimes
11	R	10/7.5	10/7.5	23.00	M	Mechanical & Industrial Engineering	Regularly
12	R	10/7.5	10/7.5	32.00	M	Mechanical & Industrial Engineering	Rarely

DOCUMENT CONTROL DATA
(Security classification of the title, body of abstract and indexing annotation must be entered when the overall document is classified)

1. ORIGINATOR (The name and address of the organization preparing the document, Organizations for whom the document was prepared, e.g. Centre sponsoring a contractor's document, or tasking agency, are entered in section 8.)

Publishing: DRDC Toronto
Performing: University of Toronto, Department of Mechanical and Industrial Engineering
Monitoring:
Contracting: DRDC Toronto

2. SECURITY CLASSIFICATION (Overall security classification of the document including special warning terms if applicable.)

UNCLASSIFIED

3. TITLE (The complete document title as indicated on the title page. Its classification is indicated by the appropriate abbreviation (S, C, R, or U) in parenthesis at the end of the title)

Developing Human−Machine Interfaces to Support Appropriate Trust and Reliance on Automated Combat Identification Systems (U)
Développement d'interfaces homme−machine pour appuyer la confiance dans les systèmes automatisés d'identification au combat Rapport d'étape pour les jalons 1, 2 et 3 (U)

4. AUTHORS (First name, middle initial and last name. If military, show rank, e.g. Maj. John E. Doe.)

G. A. Jamieson; L. Wang

5. DATE OF PUBLICATION (Month and year of publication of document.)

October 2007

6a NO. OF PAGES (Total containing information, including Annexes, Appendices, etc.)

82

6b. NO. OF REFS (Total cited in document.)

72

7. DESCRIPTIVE NOTES (The category of the document, e.g. technical report, technical note or memorandum. If appropriate, enter the type of document, e.g. interim, progress, summary, annual or final. Give the inclusive dates when a specific reporting period is covered.)

Contract Report

8. SPONSORING ACTIVITY (The names of the department project office or laboratory sponsoring the research and development − include address.)

Sponsoring: DRDC
Tasking:

9a. PROJECT OR GRANT NO. (If appropriate, the applicable research and development project or grant under which the document was written. Please specify whether project or grant.)

15au

9b. CONTRACT NO. (If appropriate, the applicable number under which the document was written.)

W7711−068000/001/TOR

10a. ORIGINATOR'S DOCUMENT NUMBER (The official document number by which the document is identified by the originating activity. This number must be unique to this document)

DRDC Toronto CR 2007−148

10b. OTHER DOCUMENT NO(s). (Any other numbers under which may be assigned this document either by the originator or by the sponsor.)

11. DOCUMENT AVAILABILITY (Any limitations on the dissemination of the document, other than those imposed by security classification.)

Unlimited distribution

12. DOCUMENT ANNOUNCEMENT (Any limitation to the bibliographic announcement of this document. This will normally correspond to the Document Availability (11), However, when further distribution (beyond the audience specified in (11) is possible, a wider announcement audience may be selected.))

Unlimited announcement

DOCUMENT CONTROL DATA
(Security classification of the title, body of abstract and indexing annotation must be entered when the overall document is classified)

13. ABSTRACT (A brief and factual summary of the document. It may also appear elsewhere in the body of the document itself. It is highly desirable that the abstract of classified documents be unclassified. Each paragraph of the abstract shall begin with an indication of the security classification of the information in the paragraph (unless the document itself is unclassified) represented as (S), (C), (R), or (U). It is not necessary to include here abstracts in both official languages unless the text is bilingual.)

(U) This research tested the effects of system reliability information and interface features on human trust and reliance on individual combat ID systems. Experiment I showed that participants had difficulty in estimating the reliability of the 'unknown' feedback from these systems. Providing the reliability information led to appropriate reliance on that feedback. Experiment II showed that participants' trust in the 'unknown' feedback was influenced by the system's activation mode and the 'unknown' feedback form, but their reliance on 'unknown' feedback was not affected. In addition, a new method was proposed to measure reliance on automation. This measure was used effectively in both experiments, and demonstrated several advantages over previous methods. Finally, implications for the design of interfaces for individual combat ID systems and the training of infantry soldiers were drawn from the results of the studies.

(U) L'objet de la recherche était de tester les effets de l'information sur la fiabilité d'un système et des fonctions d'une interface sur la confiance humaine dans les systèmes d'identification au combat. L'expérience I a démontré que les participants avaient de la difficulté à évaluer la fiabilité des commentaires de « source inconnue » relatifs à ces systèmes. Le fait de leur fournir de l'information sur la fiabilité les a menés à avoir confiance dans les commentaires, de façon appropriée. L'expérience II a démontré que la confiance des participants dans les commentaires de « source inconnue » était influencée par le mode d'activation du système et le formulaire de commentaires de « source inconnue », mais leur confiance dans les commentaires de « source inconnue » n'était pas touchée. De plus, une nouvelle méthode a été proposée pour mesurer la confiance dans l'automatisation. Cette méthode a été utilisée de façon efficace dans le cadre des deux expériences et a présenté plusieurs avantages par rapport aux méthodes antérieures. Enfin, les résultats des études ont permis d'établir les implications de la conception des interfaces des systèmes d'identification au combat et de la formation des soldats d'infanterie.

14. KEYWORDS, DESCRIPTORS or IDENTIFIERS (Technically meaningful terms or short phrases that characterize a document and could be helpful in cataloguing the document. They should be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location may also be included. If possible keywords should be selected from a published thesaurus, e.g. Thesaurus of Engineering and Scientific Terms (TEST) and that thesaurus identified. If it is not possible to select indexing terms which are Unclassified, the classification of each should be indicated as with the title.)

(U) human−machine interfaces; trust; reliance; system reliability; combat identification; automated intelligent systems; performance feedback