AFRL-AFOSR-VA-TR-2019-0104

Role of Effective Communication in Trust Building: Application to Human-Computer Interaction

Mohammad Khan
UNIVERSITY OF CONNECTICUT

03/07/2019
Final Report

Air Force Research Laboratory
AF Office Of Scientific Research (AFOSR)/ RTA2
Arlington, Virginia 22203
Air Force Materiel Command

DISTRIBUTION A: Distribution approved for public release.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
16-04-2019	Final Performance	30 Sep 2015 to 29 Sep 2018

4. TITLE AND SUBTITLE
Role of Effective Communication in Trust Building: Application to Human-Computer Interaction

5a. CONTRACT NUMBER

5b. GRANT NUMBER
FA9550-15-1-0490

5c. PROGRAM ELEMENT NUMBER
61102F

6. AUTHOR(S)
Mohammad Khan, Ross Buck

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
UNIVERSITY OF CONNECTICUT
438 WHITNEY RD EXTENSION UNIT 1133
STORRS, CT 06269-1133 US

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AF Office of Scientific Research
875 N. Randolph St. Room 3112
Arlington, VA 22203

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/AFOSR RTA2

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-AFOSR-VA-TR-2019-0104

12. DISTRIBUTION/AVAILABILITY STATEMENT
A DISTRIBUTION UNLIMITED: PB Public Release

13. SUPPLEMENTARY NOTES

14. ABSTRACT
Expressiveness and nonverbal communication play a major role in the development and maintenance of trust in humanhuman interaction. However, current designs of human-computer interactions for safety-critical systems often heavily rely on information reduction/hiding rather than revealing to reduce the cognitive load on users, and do not adapt based on user's emotions. This approach of information hiding and lack of consideration of user's emotions while interacting is likely to affect trust building negatively and instead, can foster suspicion and doubt, making it difficult to perform critical tasks with confidence. To address these limitations, this proposal investigates the challenge of designing trust-inducing humancomputer interactions taking the trust factor and user emotions explicitly into account, and investigates the following key research questions: (i) What information regarding system performance is relevant to 'trust' and need to be communicated? (ii) How to effectively communicate information regarding system states to end users to promote trust in safety-critical scenarios? and (iii) How to maintain trust over time? Our key contributions are as follows.

15. SUBJECT TERMS
Human Systems, Autonomous Systems, Human Computer Interaction

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			RIECKEN, RICHARD
Unclassified	Unclassified	Unclassified	UU		

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18

				19b. TELEPHONE NUMBER *(Include area code)* 703-941-1100

Role of Effective Communication in Trust Building: Application to Human-Computer Interaction

Contract: AFOSR FA9550-15-1-0490
PI: Mohammad Maifi Hasan Khan, University of Connecticut, maifi.khan@uconn.edu
Co-Investigators: Ross Buck, University of Connecticut; Emil Coman, University of Connecticut Health Center

Problem/Objective

Expressiveness and nonverbal communication play a major role in the development and maintenance of trust in human-human interaction. However, current designs of human-computer interactions for safety-critical systems often heavily rely on information reduction/hiding rather than revealing to reduce the cognitive load on users, and do not adapt based on user's emotions. This approach of information hiding and lack of consideration of user's emotions while interacting is likely to affect trust building negatively and instead, can foster suspicion and doubt, making it difficult to perform critical tasks with confidence. To address these limitations, this proposal investigates the challenge of designing trust-inducing human-computer interactions taking the trust factor and user emotions explicitly into account, and investigates the following key research questions: *(i) What information regarding system performance is relevant to "trust" and need to be communicated? (ii) How to effectively communicate information regarding system states to end users to promote trust in safety-critical scenarios? and (iii) How to maintain trust over time?* Our key contributions are as follows.

Results & Impact

In our work, we identified that system performance information and role (e.g., system administrator vs. operator) affect participants' reasoning differently depending on risk level [2]. Results further indicated that risk level has a significant main effect on negative individualistic and negative prosocial emotions [3]. Participants assigned to the high risk scenario anticipated more intense negative individualistic (e.g., nervous) and negative prosocial (e.g., resentful, lonely) emotions and less intense positive (e.g., happy, proud) emotions than participants assigned to the medium and low risk scenarios [3]. In our follow up study [1, 5], we investigated how different types of trust-related information about a drone system (e.g., purpose, process, performance) influence emotions anticipated as an operator. Our findings suggest that (a) users' risk taking tendencies influence trustworthiness perceptions of systems, (b) different types of information about a system have varied effects on the trustworthiness dimensions, and (c) institutions play an important role in users' calibration of trust. We also found that propensity to trust, risk-taking tendencies, and institutional trust influenced the intensity of anticipated emotions. These findings indicate that contextual risk and a user's role can influence emotions and attitudes toward safety-critical systems differently, and should be considered explicitly while designing interactions.

Next, we pose the question—*Are the people behind an automated system implicated in its mistakes? Or Is the system itself deemed a responsible actor that can repair broken trust?* While users likely correctly understand that these machines are products of human design, evidence suggests that humans respond to computers socially. Perhaps, then, the user is engaging in a trusting relationship with the system itself. If so, users may be prone to poor "trust calibration." Our study sought to elucidate the separation of system and developers by investigating how the ***attribution of blame*** for system errors influences users' trust. We recruited 147 participants on Amazon Mechanical Turk (MTurk) to play an online game (developed by us) where they collaborated with an Automated Target Detection (ATD) system in 5 rounds of an image classification task. In a 2 (reliability) x 3 (blame) between-group study, participants interacted with a high or low reliability system. After each round, the system displayed a text message acknowledging its errors in identifying images in the previous round, attributing blame either internally ("I was not able..."), pseudo-externally ("The developers were not able..."), or externally ("A third-party algorithm that I used was not able..."). Participants chose how many images to allocate to the automation and were compensated based on their combined performance with the ATD system. After gameplay, participants

responded to a survey. We found that reliability influenced both behavioral and subjective trust, while blame influenced subjective trust. Specifically, internal blame was regarded more positively than pseudo-external blame, suggesting that a system's errors are not considered the same as the developers' errors. We found a main effect of reliability on both behavioral trust and trusting perceptions. Moreover, we found that internal blame by the system and blame of the developers were perceived differently. These findings suggest that, when it comes to trust, automated systems are not treated merely as reflections of their developers, but as distinct social actors. This notion is critical for designers to ensure that users are able to accurately gauge the trustworthiness of systems, and for fostering a future of healthy human-machine relationships.

Finally, we focus on *trust repair* and investigated the effects of warning reliability and system performance feedback on trust and emotional states using a two-way 2 (warning reliability: high/low) X 2 (feedback: present/ absent) between groups in-lab study. In this in-lab study, we recruited 57 participants who played 4 rounds of a 7-minute long video game (implemented by us) simulating a drone operation. The game objective was to find and neutralize parked enemy vehicles on the street of a city. Participants received system warning messages that may help them to avoid possible system failures during the gameplay. After each warning, half of the participants were given feedback regarding whether the warning was true-alarm or false-alarm using audio message. Contrary to our hypothesis, results indicated that feedback negatively affected users' positive emotions and trust in the system, and increased negative emotions. Moreover, results indicated that hostility and loneliness emotions were higher for the feedback present groups. Regression analysis showed that the positive emotions were positively correlated with the trust factors (i.e., performance, process, and purpose) and the negative emotions (i.e., hostility and loneliness) were negatively correlated with the trust factors.

In addition to using a scale to rate emotions, we used iMotion software program to record and parse emotions from facial expressions. This software can detect and readily output micro and macro level expressions for seven primary emotions (i.e., anger, contempt, disgust, fear, joy, sadness, and surprise) and two complex emotions (i.e., frustration and confusion). Facial expression analysis revealed that feedback present groups experienced fear emotion significantly less percent of the time overall compared to the feedback absent groups. During the warning segment, feedback present groups experienced fear emotion significantly less percent of the time than the feedback absent groups as well. We observed a significant main effect of reliability on experienced frustration emotion during the warning segments. Specifically, high reliability warning groups experienced frustration emotion significantly more percent of the time than the low reliability warning groups. Furthermore, high reliability groups experienced anger emotion significantly more percent of the time than the low reliability groups. We argue that the observed negative effect of feedback on emotion and trust, while unexpected, is not necessarily a "bad" thing. Rather, it might be an effective way to nudge participants to gauge the reliability of automation systems carefully and make decisions while mindfully processing risks. As such, feedback mechanisms can facilitate calibration of trust in unreliable systems, preventing the possibility of *"overreliance"* on automation.

To summarize, our investigation found that providing feedback regarding the system performance decreases the level of trust in the system, which might be an effective way to counter possible misuse (i.e., overreliance) of automation. On the other hand, not providing feedback might be an effective way to nudge operators to trust the system more if desired. Furthermore, our findings suggest that providing feedback is likely to reduce mental workload, calling for further research in the design of feedback mechanisms as an effective way to calibrate emotion and trust. Finally, understanding how user's emotions relate to their trust in systems can not only improve system communication strategies and design for trust calibration, but can highlight the social and emotional ways in which human responds to computer interaction partners, impacting research in other domains such as human-robot interactions.

Publications/Accomplishments

1. Theodore Jensen, Yusuf Albayram, Mohammad Maifi Hasan Khan, Ross Buck, Emil Coman, and Md Abdullah Al Fahim. *Initial Trustworthiness Perceptions of a Drone System based on Performance and Process Information.* In Proceedings of the 6th International Conference on Human-Agent Interaction, pp. 229-237. ACM, 2018.
   Nominated for Best Paper Award.

2. Yusuf Albayram, Theodore Jensen, Mohammad Maifi Hasan Khan, Ross Buck, and Emil Coman. *Investigating the Effect of System Reliability, Risk, and Role on Users' Emotions and Attitudes toward a Safety-Critical Drone System.* International Journal of Human–Computer Interaction (2018): 1-12.

3. Yusuf Albayram, Mohammad Maifi Hasan Khan, Theodore Jensen, Ross Buck, and Emil Coman. *The Effects of Risk and Role on Users' Anticipated Emotions in Safety-Critical Systems.* In International Conference on Engineering Psychology and Cognitive Ergonomics, pp. 369-388. Springer, Cham, 2018.

4. Yusuf Albayram, Mohammad Maifi Hasan Khan, Ross Buck, Emil Coman. Investigating the Effect of System Reliability on Users' Emotions (Abstract). Presented as a thematic flash talk at the annual conference of the Society for Affective Science (SAS), Boston, MA, April 27-29, 2017.

5. Theodore Jensen, Mohammad Maifi Hasan Khan, Yusuf Albayram, Md Abdullah Al Fahim, Ross Buck, Emil Coman. Anticipated Emotions in Initial Trust Evaluations of a Drone System based on Performance and Process Information. Submitted to International Journal of Human–Computer Interaction, November 30, 2018 (under review).

6. Md Abdullah Al Fahim, Mohammad Maifi Hasan Khan, Yusuf Albayram, Theodore Jensen, Ross Buck, and Emil Coman. *Effect of Feedback and Warning Reliability on Trust, Emotions, and System Usage.* Submitted to ACM conference on Designing Interactive Systems (DIS), 2019 (under review).

7. Theodore Jensen, Yusuf Albayram, Mohammad Maifi Hasan Khan, Md Abdullah Al Fahim, Ross Buck, and Emil Coman. *The Apple Does Fall Far from the Tree: User Separation of a System from its Developers in Human-Automation Trust Repair.* Submitted to ACM conference on Designing Interactive Systems (DIS), 2019 (under review).

8. Md Abdullah Al Fahim, Mohammad Maifi Hasan Khan, Yusuf Albayram, Theodore Jensen, Ross Buck, and Emil Coman. *Effect of Feedback and Warning Reliability on Emotion and Task Load: An exploratory Study.* Study completed. Manuscript under preparation.

Initial Trustworthiness Perceptions of a Drone System based on Performance and Process Information

Theodore Jensen
Department of Computer Science & Engineering
University of Connecticut
Storrs, Connecticut
theodore.jensen@uconn.edu

Yusuf Albayram
Department of Computer Science & Engineering
University of Connecticut
Storrs, Connecticut
yusuf.albayram@uconn.edu

Mohammad Maifi Hasan Khan
Department of Computer Science & Engineering
University of Connecticut
Storrs, Connecticut
maifi.khan@uconn.edu

Ross Buck
Department of Communication,
University of Connecticut
Storrs, Connecticut
ross.buck@uconn.edu

Emil Coman
Health Disparities Institute,
University of Connecticut Health Center
Hartford, Connecticut
coman@uchc.edu

Md Abdullah Al Fahim
Department of Computer Science & Engineering
University of Connecticut
Storrs, Connecticut
md.fahim@uconn.edu

ABSTRACT

Prior work notes dispositional, learned, and situational aspects of trust in automation. However, no work has investigated the relative role of these factors in initial trust of an automated system. Moreover, trust in automation researchers often consider trust unidimensionally, whereas ability, integrity, and benevolence perceptions (i.e., trusting beliefs) may provide a more thorough understanding of trust dynamics. To investigate this, we recruited 163 participants on Amazon's Mechanical Turk (MTurk) and randomly assigned each to one of 4 videos describing a hypothetical drone system: one control, the others with additional system *performance* or *process*, or both types of information. Participants reported on trusting beliefs in the system, propensity to trust other people, risk-taking tendencies, and trust in the government law enforcement agency behind the system. We found that financial risk-taking tendencies influenced trusting beliefs. Also, those who received *process* information were likely to have higher integrity and ability beliefs than those not receiving *process* information, while those who received *performance* information were likely to have higher ability beliefs. Lastly, perceptions of structural assurance positively influenced all three trusting beliefs. Our findings suggest that a) users' risk-taking tendencies influence trustworthiness perceptions of systems, b) different types of information about a system have varied effects on the trustworthiness dimensions, and c) institutions play an important role in users' calibration of trust. Insights gained from this study can help design training materials and interfaces that improve user trust calibration in automated systems.

HAI '18, December 15–18, 2018, Southampton, United Kingdom

ACM ISBN 978-1-4503-5953-5/18/12…$15.00
https://doi.org/10.1145/3284432.3284435

CCS CONCEPTS

• **Human-centered computing** → *Collaborative interaction*; *HCI theory, concepts and models*; *Empirical studies in HCI*;

KEYWORDS

human-automation trust; trusting beliefs; perceived trustworthiness; initial trust

ACM Reference Format:
Theodore Jensen, Yusuf Albayram, Mohammad Maifi Hasan Khan, Ross Buck, Emil Coman, and Md Abdullah Al Fahim. 2018. Initial Trustworthiness Perceptions of a Drone System based on Performance and Process Information. In *6th International Conference on Human-Agent Interaction (HAI '18), December 15–18, 2018, Southampton, United Kingdom.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3284432.3284435

1 INTRODUCTION

Recent advancement in computing technologies has opened up the possibility of next-generation cyber-human systems (e.g., multi-UAV supervisory control, remotely operated air and ground vehicles, robot-assisted emergency response systems) where users will be required to interact and cooperate with autonomous or semi-autonomous systems to accomplish challenging and risky tasks (e.g., surveillance, battlefield operation, navigation). Prior work has noted that inappropriate levels of user "trust" in such safety-critical systems, including both *overtrust* and *undertrust*, can have undesirable and even fatal consequences [35], calling for further research on how to achieve appropriate "trust calibration." Characteristics of the human operator, the environment, and the automated system have been identified as factors that can influence trust in an automated or autonomous trustee. Hoff and Bashir suggested a three-layered model in which these factors contribute to *dispositional*, *situational*, and *learned* trust, respectively [19].

Interestingly, while a significant amount of work has investigated these aspects of trust in the context of human-machine interactions, very little is known regarding "initial trust" in this context. Initial trust, or trust in an unfamiliar party [30], contributes to the "risk-taking in relationship" that defines later trust [2, 27]. Moreover, first

impressions and system training have been found to affect later trust and reliance in safety-critical systems [6, 13].

To complement prior efforts focusing on trust during human-machine interactions, the present study focuses on factors that contribute to the development of initial trust. Specifically, we aim to observe how characteristics of the operator (i.e., dispositional), perceptions of the institution behind a system (i.e., situational), and information about a system (i.e., learned) influence initial trustworthiness perceptions of that system. Toward this, we designed 4 narrated videos with different information about a hypothetical drone system. We recruited 163 naive participants (i.e., without prior experience with drone systems) on Amazon MTurk, randomly assigned each to one of the 4 video groups, and asked them to imagine that they were going to operate the drone in a safety-critical task. Given the sample's lack of exposure to similar technologies, we can observe how system information contributes to learned trust, and how dispositional and situational factors simultaneously influence trust evaluations.

After watching their video, participants rated statements on trusting beliefs in the system's ability, integrity, and benevolence, and institutional trusting beliefs in situational normality and structural assurance. They also reported on demographics, propensity to trust other people, and risk-taking tendencies. We discuss the implications of our findings for human-computer trust research and system design.

2 RELATED WORK

2.1 Trust

The concepts of "trust" and "trustworthiness" have been extensively studied by various research communities such as psychologists and communication theorists in the context of human-human interactions. While some refer to it as a behavioral state, an individual disposition, a set of social expectations, or an emotion, there is consensus that risk lies at the heart of trust [1, 5, 9, 38]. Mayer et al.'s model gives one of the most widely accepted definitions of human-human trust: "The willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that party" [27].

In this relationship between trustor and trustee, trust is affected by characteristics of the trustor, the trustee, and the situational context [27]. A trustor's trusting *beliefs* contribute to their trust, realized as risk-taking in the relationship with the trustee. These beliefs are perceptions of a trustee's trustworthiness, which Mayer et al. operationalize into three characteristics: *ability*, *integrity*, and *benevolence*. *Ability* relates to the trustee's skills or competencies within some domain. *Integrity* reflects that the trustee "adheres to a set of principles" that is acceptable to the trustor. *Benevolence* relates the extent to which a trustee will do good for the trustor [27].

2.2 "Trust" in Human-Machine Interactions

Many researchers have applied insights from the human-human trust literature to study trust between a human and a machine, computer, or automated system. This notion is motivated by work in the "Computers are Social Actors" paradigm, which has shown that users apply various social norms to computer interaction partners. These include perceptions of praise and derogation, treating different voices as distinct social actors, applying gender stereotypes, and using politeness [34]. In the latter study, participants worked with a computer and subsequently assessed its performance in one of three ways: on the same computer, with pencil and paper, or on a different computer [37]. Those in the same computer condition rated the computer as more friendly and competent than those in the other conditions, suggesting that politeness was employed when giving a direct assessment. Despite these responses, at the end of the experiment, participants denied acting polite toward the computer. Reeves and Nass suggest that these social responses occur "mindlessly" in that, even though they may think it's nonsensical, people apply the notion, "When in doubt, treat it as human," when interacting with computers [33, 37].

Among prior efforts that attempt to understand trust in the human-machine context, Bonnie Muir defined trust in human-machine interaction as a function of expectations held by a member of a system regarding the persistence, technical competence, and fiduciary responsibilities from another member of the system [32]. More recently, in line with Mayer et al.'s human-human model, Hoff and Bashir noted characteristics of the human operator, the environment, and the automated system as factors influencing trust in an automated trustee. They suggest a three-layered model in which these factors contribute to *dispositional*, *situational*, and *learned* trust, respectively [19].

Hoff and Bashir suggest that age, gender, and personality differences are components of *dispositional* trust [19], while prior work has begun to investigate the factors influencing dispositional trust in automation [20].

Prior work has further identified situational normality and structural assurance as two aspects of institutional trust, where situational normality consists of beliefs about an institution's trustworthiness in a given context [25, 29], and structural assurance relates to a belief that guarantees or "safety nets" are in place to protect the user if something goes wrong [25, 29]. These institutional perceptions constitute one aspect of Hoff and Bashir's *situational* trust.

Regarding *learned* trust (i.e., in a specific trustee), Lee and Moray refer to performance, process, and purpose as the bases for trust in automation [23]. Performance refers to the consistency and reliability of system behavior and its history of operation. Process relates to how the system operates and the degree to which its algorithms are appropriate in a given context. Purpose is the extent to which the system is being used within the "realm of the designer's intent" [23]. The lexical dissonance between Mayer et al.'s perceived trustworthiness characteristics and Lee and Moray's terms distinguishes human-human from human-computer trust. In general, a trustor is concerned with the what (*ability* or *performance*), the how (*integrity* or *process*), and the why (*benevolence* or *purpose*) of their interaction partner's behavior. Trust "calibration" occurs when a system user adjusts their perceptions of these characteristics to better fit a system's actual reliability [24]. While prior work on trust in e-Commerce and online recommendation agents [2, 22, 29–31] have operationalized perceived trustworthiness using ability, integrity, and benevolence items adapted from Mayer's research [26], to the best of our knowledge this approach has not been taken in the trust in automation literature.

2.3 Initial Trust and the Current Study

"Initial trust," or trust in an unfamiliar party [30] has been shown to precede and contribute to the risk-taking in relationship that defines later trust [2, 27]. First impressions and system training have been found to affect later trust and reliance in safety-critical systems as well [6, 13]. Furthermore, research on organizational information systems and e-Commerce has found factors such as reputation and perceived usefulness to influence initial trust [2, 22, 25, 30]. Despite the importance of initial trust, prior work on safety-critical systems has not explored this and generally observes trust during interaction [3, 10].

As such, to complement prior efforts, we aim to observe how characteristics of the operator, perceptions of the law enforcement agency behind the system, and information about a drone system influence initial perceived trustworthiness. We expect that, in addition to general propensity to trust others [39], individual differences in risk-taking are influential in trust evaluations of safety-critical systems. Furthermore, by measuring "institutional trust," we treat the system itself as a distinct trustee, and avoid the assumption that trust in the system is the same construct as trust in system designers or other members of the organization.

A main goal of this study is to observe whether users perceive an automated system to have ability, integrity, and benevolence, and the extent to which the aforementioned *dispositional*, *situational*, and *learned* factors contribute to these perceptions. The details of our study are presented in the following sections.

3 METHODOLOGY

The purpose of this study is to observe how information about a system's performance and process can be communicated to influence initial beliefs in the ability, integrity, and benevolence of a drone system, and how personal and institutional factors affect these perceptions. More specifically, this study seeks to answer the following research questions:

- **RQ1:** Does a person's propensity to trust other humans translate to a computer trustee?
- **RQ2:** Do a person's risk-taking tendencies influence their initial trust?
- **RQ3:** How do performance and process information influence beliefs in a system's ability, integrity, and benevolence?
- **RQ4:** What role does institutional trust play in trust of a computer system itself?

3.1 Design of Videos

To investigate the aforementioned research questions, we created 4 videos describing a hypothetical drone system.

We chose video as the mode of communication because there is evidence that videos can be effective in introducing a system to users (e.g., system training) by utilizing both visual and auditory information processing channels, which leads to higher engagement [7, 17, 28, 36, 41]. This method also allows us to isolate the effects of information participants receive from factors such as experience and interface features that can influence trustworthiness perceptions during use. We chose to investigate a drone system because most individuals are familiar with their safety-critical applications.

Label	n	Link to Video	Length
Control	41	https://youtu.be/DuMwSsrEG5s	50 s
Performance	39	https://youtu.be/RJdwtSuGmAc	71 s
Process	39	https://youtu.be/2BTbNTAG19A	70 s
Perf-Proc	44	https://youtu.be/c5JrIdQNkY4	92 s

Table 1: List of the 4 videos used in the study, which can be viewed on YouTube. The original video can be found at: https://www.dvidshub.net/video/411919/ mq1b-predator-gcs-broll.

Information was given in the form of narration. The content was reviewed and revised by authors over several iterations to ensure clarity for our naive participants and relevance to Lee and Moray's definitions [23] of *performance* and *process* applied to our drone system. These two information types were defined as follows:

- ***Performance*** refers to the consistency and reliability of system behavior. The narrated performance information included how external/internal factors (e.g., poor network connections, software glitches) can influence the reliability of the system and what the consequences might be.
- ***Process*** details the qualities that govern system behavior, such as its algorithms. The narrated process information included how the system behaves to make it robust against possible failures (e.g., sensors are used to monitor flight stability).

Lee and Moray's third characteristic, the *purpose* of the system, was given in all videos. We determined this was necessary to give participants sufficient information to understand the system and their role as operator. Thus, the control group watched a video containing only "baseline" information (i.e., describing what the system was used for), and the three experimental groups watched a video containing the same baseline content followed by either performance or process information, or both. Table 2 shows the full narration transcript.

The video's visual content was taken from a publicly available video of drone operation. The original audio was replaced by narration recorded by one of the researchers. To investigate the effect of performance and process information alone, as well as their interaction, videos were trimmed to the length of their narration. This avoided both video playing without narration and repeated visual content. Longer videos therefore contain visual content that shorter videos do not. Because the video displays neutral images of operators at a control panel, we expect that the narration describing a safety-critical drone task was more salient. However, we acknowledge the potential effect of the visual content and refer the reader to Table 1 to view the videos on YouTube. While participants watched clear versions of the videos, we blurred out certain parts in the shared links for privacy reasons.

3.2 Recruitment

We posted the study as a Human Intelligence Task (HIT) on Amazon's Mechanical Turk (MTurk) service. Using MTurk's eligibility criteria, the HIT was made available to users 18 years or older, living in the United States, with at least 1000 completed HITs and a 95% HIT approval rating. When participants accepted the HIT,

Baseline	Hello! The video you are watching presents a hypothetical scenario where operators are using a drone system to assist government law enforcement in stopping human traffickers. The system consists of an Unmanned Aerial Vehicle, or UAV, and a display that the operator observes while controlling the system. The operator is responsible for navigation of the drone and reporting locations of suspected human traffickers. Timely and accurate identification of violent criminals is extremely important as failures can put innocent civilians' and law enforcement officers' lives in danger. While the operators may shoot at targets if necessary, this is used only as the last option, as it could lead to hitting innocent civilians near the target or causing property damage.
Performance	While the system operates effectively most of the time, there can be occasional errors that impact video quality and drone maneuverability, caused by factors such as poor network connections and software glitches. As a result, operators may experience rare events such as screen blackouts or loss of connectivity lasting at most a few seconds.
Process	To make the system robust against such failures, the UAV has on-board algorithms that use sensors to improve flight stability and maneuverability. Information about system health is also automatically monitored and sent back to the operator over a network connection. This allows the operator to monitor and override system control if needed.

Table 2: Narration script for the videos. All videos contained "baseline" information describing the purpose of the system, while the *Performance* video additionally contained the performance information, *Process* the process information, and *Perf-Proc* both types of additional system information.

they were shown an information sheet and link connecting them to the study hosted on our university's Qualtrics server.

There were 3 pre-screening questions to prevent participants from guessing the eligibility criterion. Participants had to answer "No" to the question *"Have you ever operated drones in the past?"*. This screened out individuals having operated recreational drones in addition to systems like that described in our study, ensuring that we observed *initial* trust. We did not disclose this eligibility criterion to any participant.

Ineligible participants were informed that they could not participate or be compensated. Eligible participants took the survey. At the end, these participants were given a code generated on Qualtrics to submit to MTurk for $3 of compensation. On average, the survey took participants 14.3 minutes (Median = 12.4 minutes, SD = 8.5 minutes).

The study was approved by our university's Institutional Review Board.

3.3 Survey Design

In the survey, participants first answered demographic questions on age, gender, computer proficiency, race, education, and military experience. Next, each viewed their randomly assigned video and was subsequently shown the following text:

Now imagine that you are working for a law enforcement agency as the operator of the presented drone system. Your task is to identify, track, and neutralize the vehicles of human traffickers who could harm civilians if not detained. Please note that failure to identify violent criminals such as human traffickers can put innocent civilians' and law enforcement officers' lives at danger. Please answer the following questions assuming the presented operating conditions.

Participants were then asked to reiterate the scenario in their own words to ensure that they understood their task and role as operator.

Next, we evaluated Mayer et al.'s trusting beliefs in the system's *ability*, *integrity*, and *benevolence* [27]. These items were adapted from [26, 30] to refer to our drone system (see Table 3).

Subsequently, participants answered questions about situational normality and structural assurance. These institutional trust items were adapted from [25, 29] to refer to the government law enforcement agency in our hypothetical scenario.

Participants then reported on their propensity to trust other people using a 12-item scale [12] and risk-taking tendencies in 5 domains (financial, ethical, health/safety, recreational, social) using the 30-item domain-specific risk-taking (DOSPERT) scale [4, 18, 40, 43].

We included two questions in the survey asking for a specific answer (e.g., "Mostly Agree") to identify inattentive participants. We also included two manipulation check items at the end of the survey to validate that the system's performance and process were communicated in the video.

4 EVALUATION

4.1 Sample Demographics

Of the 200 participants eligible after pre-screening, we removed the data of those who incorrectly answered at least one multiple choice attention check question, entered an ineligible age, or misunderstood the scenario based on their post-video reiteration. For the latter, participants who gave an unrelated response, referred to the "operator" in the third-person (i.e., suggesting that they were not imagining being the operator) or mentioned something not expressed in the video (e.g., "the military," "drug sales," "child sex traffickers") were removed from the data. Lastly, to ensure the video was fresh in participants' minds, we removed data of those who waited greater than 10 minutes after their video ended to advance to the next part of the survey. Ultimately, 163 were retained for analysis and balanced among video groups (see Table 1).

The sample consisted of 89 (54.6%) male and 74 (45.4%) female participants with ages ranging from 20 to 64 (Mean = 35.3, SD = 10.0). When asked about computer proficiency, 58 (35.6%) participants reported being "Competent," 82 (50.3%) "Proficient," and 23 (14.1%) "Expert." There were 123 white/Caucasian (75.5%), 18 African American (11.0%), 6 Hispanic (3.7%), 11 Asian (6.7%), 2

Trusting Belief	Items
Ability	- The drone system would be competent and effective at assisting in tracking enemy targets. - The drone system would perform its role of neutralizing enemy targets very well. - Overall, the drone system would be a capable and proficient means for stopping the targets. - In general, the drone system would be very knowledgeable about stopping criminals.
Integrity	- The drone system would be truthful in its communication with me. - I would characterize the drone system as honest. - The drone system would keep its commitments. - The drone system would be sincere and genuine. - The drone system would perform as expected.
Benevolence	- I believe that the drone system would operate in my best interest. - If I required help, the drone system would do its best to help me. - The drone system would be concerned about my well-being, not just its own. - The drone system would be concerned about the well-being of officers on the ground. - The drone system would be concerned about the well-being of civilians.

Table 3: The three trusting beliefs and their subsisting items, scored on a 7-point Likert scale from 1 = "Strongly Disagree" to 7 = "Strongly Agree." These items were adapted from [26, 30].

Native American and 3 other participants. Furthermore, 82.2% of participants reported having some post-secondary education at a college or university and 7 (4.3%) reported having served in the military.

Testing for demographic differences between video groups, we found no significant differences in terms of gender ($\chi^2(3) = 2.19$, $p = .53$). Moreover, Fisher's Exact Test revealed neither significant differences in terms of race ($p = .51$) nor military service ($p = .32$). Kruskal-Wallis tests demonstrated that groups were not significantly different in terms of age ($H(3) = 1.73$, $p = .63$), education level ($H(3) = 0.16$, $p = .98$), or computer proficiency ($H(3) = 2.60$, $p = .46$). Based on these results, we concluded that the four groups recruited were similar in terms of their demographics.

4.2 Validation of Information Types Communicated in the Videos

First, to verify whether performance and process information were communicated in the narration, we included two manipulation check statements at the end of the survey:

1. I was made aware of the drone system's performance (i.e., how effective the system is about accomplishing its goal).
2. I was made aware of the drone system's process (i.e., how the system works to accomplish its goal).

Participants rated these two items on a 7-point Likert scale from "Strongly Disagree" to "Strongly Agree." We use Mann-Whitney U-tests to compare between participants who received or did not receive a given type of information. We also report the effect size of U-tests using $r = Z/\sqrt{N}$ metric [11].

Participants who received *performance* information in their video (i.e., *Performance* and *Perf-Proc* groups) rated their awareness of the drone system's performance higher than other participants (i.e., *Control* and *Process* groups), though this difference was only marginally significant ($U = 2841.00$; $p = .10$; $r = -.13$). It may be that because the *performance* information mentioned potential system errors, these participants actually felt somewhat unaware of the system's performance.

Participants who received *process* information in their video (i.e., *Process* and *Perf-Proc* groups) rated their awareness of the system's process significantly higher than other participants (i.e., *Control* and *Performance* groups) ($U = 2667.50$; $p = .02$; $r = -.18$).

4.3 Building Multiple Linear Regression Models for Perceived Ability, Integrity, and Benevolence

To answer our research questions, we constructed multiple linear regression models that predict participants' initial trustworthiness perceptions in the system's ability, integrity, and benevolence from a set of input predictors.

Before building models, we verified the reliability of our independent and dependent scales using Cronbach's α. The 12-item propensity to trust scale had excellent reliability ($\alpha = .95$). Each of the DOSPERT 6-item risk domain scales had at least acceptable reliability (ethical $\alpha = .80$; financial $\alpha = .81$; health/safety $\alpha = .71$; recreational $\alpha = .81$; social $\alpha = .74$). For institutional trust, each 3-item situational normality sub-scale had good reliability (ability $\alpha = .88$; integrity $\alpha = .94$; benevolence $\alpha = .87$), and the 4-item structural assurance scale had excellent reliability ($\alpha = .95$). The three trusting belief scales had good reliability (4-item, ability $\alpha = .81$; 5-item, integrity $\alpha = .88$; 5-item, benevolence $\alpha = .85$). Overall, both our independent and dependent variables demonstrated acceptable reliability [14].

Next, we considered predicting each perceived trustworthiness characteristic with a different set of independent variables. For each model, we examined the Bayesian Information Criterion (BIC), a goodness-of-fit metric which also takes into account the model's complexity. A lower relative value suggests better fit for a given model.

We first included only the experimental manipulations as predictors, using *performance* and *process* variables to indicate whether each information type was presented in a participant's video. For example, the regression coefficient for the *performance* variable compares trustworthiness ratings given by participants for whom

	Ability			Integrity			Benevolence		
Predictors	β	*SE*	*p*	β	*SE*	*p*	β	*SE*	*p*
Individual Differences									
Trusting propensity	-0.16	0.08	**0.04**	-0.02	0.12	0.90	-0.15	0.13	0.27
Risk-taking:									
Ethical	-0.02	0.07	0.73	0.09	0.10	0.41	0.18	0.12	0.14
Financial	0.13	0.05	**0.02**	0.23	0.08	**<0.01**	0.32	0.09	**<0.01**
Health/safety	-0.12	0.07	0.09	0.00	0.10	0.99	-0.08	0.12	0.50
Recreational	-0.01	0.06	0.83	-0.13	0.09	0.12	-0.04	0.10	0.71
Social	0.05	0.06	0.35	-0.09	0.08	0.30	-0.15	0.10	0.12
System Information									
Perf	0.34	0.17	**0.04**	0.29	0.24	0.23	0.35	0.28	0.22
Proc	0.37	0.17	**0.03**	0.55	0.25	**0.03**	0.57	0.29	0.05
Perf x Proc	-0.36	0.23	0.13	-0.57	0.34	0.10	-0.53	0.40	0.18
Institutional Trust									
Structural Assurance	0.38	0.11	**<0.01**	0.45	0.16	**<0.01**	0.42	0.19	**0.03**
Situational Normality:									
Ability	0.12	0.12	0.32	0.20	0.18	0.26	0.20	0.21	0.33
Integrity	0.13	0.11	0.23	-0.21	0.16	0.19	-0.25	0.18	0.18
Benevolence	-0.09	0.12	0.42	0.13	0.17	0.44	0.19	0.20	0.36
Constant	2.65	0.45	**<0.01**	1.81	0.65	**<0.01**	1.29	0.76	0.09
Adjusted R^2	0.4888			0.3254			0.2623		
F-value	$F(13, 149) = 12.92$ ($p < .001$)			$F(13, 149) = 7.01$ ($p < .001$)			$F(13, 149) = 5.43$ ($p < .001$)		

Table 4: Results of the three separate multiple linear regressions, each predicting a trusting belief (ability, integrity, or benevolence) based on the various predictors. *p*-values which are significant at the 0.05 level are shown in bold.

performance information was present (i.e., *Performance* and *Perf-Proc* groups) to those for whom it was absent (i.e., *Control* and *Process* groups). The BIC value was 487.79 for the ability regression, 567.81 for integrity, and 601.51 for benevolence. Next, we added our other independent variables to each model to observe whether there was improvement in predicting perceived trustworthiness. Specifically, each trusting belief was regressed on dispositional factors of the participant (i.e., trusting propensity, risk-taking domains), situational factors that captured participants' perceptions of the hypothetical law enforcement agency behind the system (i.e., situational normality, structural assurance), and the aforementioned learned factors (i.e., system information provided in the videos). The BIC value was 418.45 for ability, 543.25 for integrity, and 592.10 for benevolence, indicating that the fit of the models improved with the addition of dispositional and situational factors. The results of these final regressions are shown in Table 4.

4.3.1 *Dispositional Trust.*

Regarding **RQ1**, participants with a greater propensity to trust other people were more likely to rate the system's ability lower (β = -0.16, p <.05). This suggests that individuals who are more trusting of other people have less faith in the ability of the drone system. We return to the influence of trusting propensity below when reporting on institutional (i.e., situational) trust.

For **RQ2**, individuals who reported greater financial risk-taking tendencies were likely to give higher ratings for the system's ability (β = 0.13, p <.05), integrity (β = 0.23, p <.01), and benevolence (β = 0.32, p <.01). These findings support those in [20] regarding both the role of individual differences in dispositional trust, as well as the distinction between dispositional trust in humans and in automation.

4.3.2 *Learned Trust.*

Concerning **RQ3**, performance and process information given to participants appeared to influence ability and integrity beliefs.

Interestingly, despite the mention of potential system errors, participants receiving *performance* information were likely to rate the system's ability higher than those not receiving performance information (β = 0.34, p <.05). While this suggests that transparency increases perceptions of trustworthiness, the fact that users knew about errors and regarded the system as having higher ability is a poor sign for trust calibration. System designers should be explicit about their system's weaknesses, not necessarily aiming to increase trust but to promote appropriate calibration. Moreover, while our finding suggests a connection between performance information and increased ability beliefs, it may be that perceptions of system ability as measured in this study do not necessarily lead to increased behavioral trust (i.e., greater reliance on the system). Information about system errors may actually make participants less willing to use the system in certain situations despite higher trusting belief ratings.

Process information appeared to lead to increased perceptions of ability (β = 0.37, p <.05) and integrity (β = 0.55, p <.05). Again, this lends to the idea that people appreciate transparency—knowledge

about the system's underlying technologies seems to have led participants to more positively regard its ability and integrity.

Despite these findings, the interaction between *performance* and *process* information did not significantly influence any of the trusting beliefs. This suggests that a larger volume of information does not necessarily contribute positively to trusting beliefs.

4.3.3 Situational Trust.
For **RQ4**, situational normality ratings did not have a significant effect on perceived trustworthiness. However, participants who gave higher structural assurance ratings were likely to give higher ratings for the ability (β = 0.38, p <.01), integrity (β = 0.45, p <.01), and benevolence (β = 0.42, p <.01) of the drone system. This finding demonstrates the important distinction between perceptions of the system itself and of the institution behind the system. While we find that positive beliefs about institutional safeguards contributed to more positive perceptions of the system's trustworthiness characteristics, the contributions from various other factors in our model show that institutional trust is not the full picture of trust in the system.

To explore structural assurance more in depth, we ran a multiple linear regression with structural assurance as the dependent variable and the same set of independent variables (excluding institutional trust items). A significant regression equation was found ($F(9, 153) = 2.86$, p <.01) with an Adjusted R^2 of 0.0936. The only significant independent factor was propensity to trust, where more trusting participants were likely to give higher structural assurance ratings ($\beta = 0.49$, p <.01). This effect on structural assurance suggests that participants' interpersonal trust indirectly affected trustworthiness perceptions of the drone system through their perception of the law enforcement agency.

5 DISCUSSION

Our findings suggest that users' individual differences, information about a system, and institutional perceptions each contribute differently to initial beliefs about a system's ability, integrity, and benevolence. The key findings and limitations of our study are discussed below.

5.1 Effect of Risk-Taking Tendencies on Ability, Integrity, and Benevolence Beliefs

We found that individuals who reported being more likely to take financial risks rated the system's ability, integrity, and benevolence more highly. Interestingly, the largest effect (i.e., greatest regression coefficient) was on the benevolence belief, followed by integrity and, subsequently, ability. This finding lends support to our application of Mayer et al.'s trustworthiness characteristics to an automated trustee, in that each was influenced differently by financial risk-taking tendencies. For instance, the fact that benevolence beliefs were the most influenced by dispositional characteristics suggests that users did build some idea of whether the hypothetical drone system cared about them, based partly on their willingness to take risks (see "Benevolence" items in Table 3), even if it may seem inappropriate to think of a computer system as having the quality of benevolence. While this benevolence perception may reflect participants' thoughts about the institution's intentions, further investigating this phenomenon by applying the "Computers are Social Actors" paradigm could elucidate the nature of human perceptions of automated trustees. We encourage future work to further explore this relationship between individual differences and the trustworthiness characteristics to inform more user-centered designs for trust calibration.

5.2 Effect of System Information on Ability and Integrity Beliefs

Our regression also revealed that *performance* or *process* information in the narration was likely to lead to increased ability beliefs. Moreover, *process* information was likely to lead to increased integrity beliefs. We note that although our manipulation check revealed that those who received performance information reported being more aware of the system's performance than those who did not receive performance information, the difference was not significant. Nonetheless, the *performance* and *process* information (see Table 2) appear to have offered transparency, casting the system as less of a fault-prone black box. Given *performance* information, the system may have appeared more competent because its shortcomings were directly acknowledged. For *process* information, the description of the system's underlying technologies (i.e., flight stability algorithms, network connection to the operator) may have contributed to these increased perceptions of trustworthiness.

In a prior study of an autonomous driving system, Koo et al. observed that a "how" explanation (i.e., reflecting the system's process) led to more drivers drifting out of their lane [21]. Their "how" explanation, "The car is braking," may have caused reliance on the system in inappropriate circumstances (i.e., overtrust and misuse [35]) by insufficiently communicating limitations of the system to its user. While our *process* information also increased participants' perceptions of system ability and integrity, an important question is how these perceptions ultimately impact users' behaviors with the system, though this is outside of our focus on initial trust. We encourage future work to investigate how initial trust influences long-term interactions and trust calibration with safety-critical systems.

In another similar study, drivers who were shown performance information about an autonomous driving system were found to have lower trust levels than those not shown the information, and were more prepared to take over control from the automation [15]. This suggests that the information allowed for trust calibration. Likewise, take-over requests and alerts have been found to lead to safer usage and greater trust [16, 42]. While our information *increased* users' perceptions of system trustworthiness, a more appropriate goal for system designers and institutions is to give information that *improves* trust. In some cases, this may mean giving trust-reducing information that allows users to calibrate to a particular system's shortcomings. While our *performance* information could be seen as trust-reducing, those who received this information rated the system's ability higher than those who did not receive it. This is likely a result of ambiguity regarding the extent to which system errors would interfere with operation of the system in our task context.

While aforementioned prior work manipulates trust during interaction with the system, we observe users' initial impressions. Although this makes it more difficult to compare our results to

prior efforts, it allows us to isolate the effect of system information on learned trust, as participants were not affected by use-related factors such as interface design and observed system performance. We encourage future work to explore the specific role of trusting beliefs in users' reliance decisions. For example, it may be that beliefs about integrity are less critical to reliance on a system than those about its ability.

The insignificant effect of the interaction between *performance* and *process* information suggests that the mention of system errors alongside a description of the system's underlying technologies tempered trustworthiness ratings compared to either information type given on its own. In this way, trust calibration may essentially be a balancing act, as suggested in de Visser et al.'s model involving trust repair acts and trust dampening acts [8]. Safety-critical systems can incorporate communication modules that relay information to users following both system malfunctions and successful operations, in order to mitigate large changes in a user's perceptions of system trustworthiness and ensure that appropriately calibrated trust is maintained.

5.3 Effect of Institutional Trust on Ability, Integrity, and Benevolence Beliefs

We found that participants who felt more structurally assured were likely to have higher initial beliefs in the system's ability, integrity, and benevolence. We also found that trusting propensity contributed to greater perceptions of structural assurance. Although this implies that trusting individuals will be more trusting of a safety-critical system, overtrust is not desirable and can lead to use of the system in inappropriate, dangerous circumstances. Institutions that oversee safety-critical operations must carefully create structural assurances such that users are held appropriately accountable for misuse of the system (i.e., to avoid overtrust), but not necessarily for failures caused by the system itself.

5.4 Limitations

Though the present study sheds light on factors in initial trust of safety-critical systems, the reported findings should be interpreted with caution due to the following limitations of our study.

First, participants had to imagine being the operator without actually interacting with a system. Because of this, their perception of risk may have been diminished. The study also uses self-reported trust ratings which, although indicative of users' initial perceptions, were not connected to a behavioral measure. Nonetheless, our study does confirm the effect of different factors on the three perceived trustworthiness characteristics, which was the main objective of our work. Further research is needed to confirm our findings and measure the effect of initial trusting beliefs on actual system interaction over time in controlled lab settings, possibly using experimental deception to create greater perceived risk. Future work should also look at the influence of system information on trust formation in systems with different reliability levels, since trust has often been found to vary with respect to system reliability [19], and appropriate calibration of trust is our main goal.

Second, since this was an online survey, we could not guarantee that participants remained attentive during the video. It is possible that some MTurk participants were multitasking and did not pay attention to the narrated information. However, to minimize this possibility, each participant had to wait at least their video duration before proceeding and we removed data of participants whose open-ended responses demonstrated a lack of understanding of the task.

Finally, there are some limitations to our manipulations. Due to differences in narrated content, the videos were not exactly the same length. For one, this caused longer videos to contain more visual content than the shorter videos. To minimize any possible effect of visual content, we carefully chose the shots of the drone control panel and operators to be relatively neutral, especially compared to the safety-critical task described in the narration. Moreover, receiving a greater amount of information could have contributed to participants' differences in perception. However, we tested for correlation between video length and trusting belief ratings and found no significant correlations, suggesting that informational content is what influenced perceived trustworthiness.

We also acknowledge potential unintended effects of message content. In particular, the *performance* information discussed "occasional errors" and may have been perceived more negatively than the *process* information. Studying the effects of valence of system information (e.g., positive performance information vs. negative performance information) may give further insights. Likewise, interpretation of the phrase "such failures" in the *process* information may have been interpreted differently depending on whether or not a participant received *performance* information. Our manipulation check did suggest that awareness of system's performance and process was increased by the narrated information types. However, we encourage future research to build upon these initial findings to better understand various informational contributors to trustworthiness perceptions.

6 CONCLUSION

In this study, we investigated the role of dispositional, learned, and situational factors on initial trustworthiness perceptions. We designed a control video with information about a hypothetical drone system and three videos with either additional performance or process information, or both, to observe how 163 naive participants on Amazon's MTurk formed trusting beliefs based on the video. We found that risk-taking individual differences affected all three beliefs. Moreover, *performance* and *process* information were likely to increase beliefs in the system's ability, while *process* information also lead to increased integrity beliefs. We stress that increased trust is not necessarily desirable–users should instead be able to calibrate their trust in a system to an appropriate level. Lastly, we found that structural assurance was associated with increased ability, integrity, and benevolence perceptions. We believe that insights gained from this study enhance our understanding of user's multidimensional perceptions of the "trustworthiness" of safety-critical systems, which can lead to more favorable trust-related outcomes of human-computer collaborations.

ACKNOWLEDGMENTS

This material is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-15-1-0490.

REFERENCES

[1] Bernard Barber. 1983. The logic and limits of trust. (1983).

[2] Izak Benbasat and Weiquan Wang. 2005. Trust in and adoption of online recommendation agents. *Journal of the association for information systems* 6, 3 (2005), 4.

[3] David P Biros, Mark Daly, and Gregg Gunsch. 2004. The influence of task load and automation trust on deception detection. *Group Decision and Negotiation* 13, 2 (2004), 173–189.

[4] Ann-Renée Blais and Elke U Weber. 2006. A domain-specific risk-taking (DOSPERT) scale for adult populations. (2006).

[5] Ross Buck. 2014. *Emotion: A biosocial synthesis.* Cambridge University Press.

[6] Michelle S Carlson, Munjal Desai, Jill L Drury, Hyangshim Kwak, and Holly A Yanco. 2014. Identifying factors that influence trust. In *2014 AAAI Spring Symposium Series.*

[7] James M Clark and Allan Paivio. 1991. Dual coding theory and education. *Educational psychology review* 3, 3 (1991), 149–210.

[8] Ewart J de Visser, Richard Pak, and Tyler H Shaw. 2018. From "automation" to "autonomy": The importance of trust repair in human-machine interaction. *Ergonomics* just-accepted (2018), 1–33.

[9] Morton Deutsch. 1960. Trust, trustworthiness, and the F scale. *The Journal of Abnormal and Social Psychology* 61, 1 (1960), 138.

[10] Stephen R Dixon, Christopher D Wickens, and Dervon Chang. 2004. Unmanned aerial vehicle flight control: False alarms versus misses. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 48. SAGE Publications Sage CA: Los Angeles, CA, 152–156.

[11] Andy Field. 2013. *Discovering statistics using IBM SPSS statistics.* Sage.

[12] M Lance Frazier, Paul D Johnson, and Stav Fainshmidt. 2013. Development and validation of a propensity to trust scale. *Journal of Trust Research* 3, 2 (2013), 76–97.

[13] Amos Freedy, Ewart de Visser, Gershon Weltman, and Nicole Coeyman. 2007. Measurement of trust in human-robot collaboration. In *Collaborative Technologies and Systems, 2007. CTS 2007. International Symposium on.* IEEE, 106–114.

[14] Joseph A Gliem and Rosemary R Gliem. 2003. Calculating, interpreting, and reporting Cronbach's alpha reliability coefficient for Likert-type scales. Midwest Research-to-Practice Conference in Adult, Continuing, and Community Education.

[15] Tove Helldin, Göran Falkman, Maria Riveiro, and Staffan Davidsson. 2013. Presenting system uncertainty in automotive UIs for supporting trust calibration in autonomous driving. In *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications.* ACM, 210–217.

[16] Sebastian Hergeth, Lutz Lorenz, Josef F Krems, and Lars Toenert. 2015. Effects of take-over requests and cultural background on automation trust in highly automated driving. In *8th International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design.*

[17] Carol Herron, Holly York, Cathleen Corrie, and Steven P Cole. 2006. A comparison study of the effects of a story-based video instructional package versus a text-based instructional package in the intermediate-level foreign language classroom. *Calico Journal* (2006), 281–307.

[18] Scott Highhouse, Christopher D Nye, Don C Zhang, and Thaddeus B Rada. 2017. Structure of the Dospert: Is There Evidence for a General Risk Factor? *Journal of Behavioral Decision Making* 30, 2 (2017), 400–406.

[19] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors* 57, 3 (2015), 407–434.

[20] Hsiao-Ying Huang and Masooda Bashir. 2017. Personal Influences on Dynamic Trust Formation in Human-Agent Interaction. In *Proceedings of the 5th International Conference on Human Agent Interaction.* ACM, 233–243.

[21] Jeamin Koo, Jungsuk Kwac, Wendy Ju, Martin Steinert, Larry Leifer, and Clifford Nass. 2015. Why did my car just do that? Explaining semi-autonomous driving actions to improve driver understanding, trust, and performance. *International Journal on Interactive Design and Manufacturing (IJIDeM)* 9, 4 (2015), 269–275.

[22] Marios Koufaris and William Hampton-Sosa. 2004. The development of initial trust in an online company by new customers. *Information & management* 41, 3 (2004), 377–397.

[23] John Lee and Neville Moray. 1992. Trust, control strategies and allocation of function in human-machine systems. *Ergonomics* 35, 10 (1992), 1243–1270.

[24] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[25] Xin Li, Traci J Hess, and Joseph S Valacich. 2008. Why do we trust new technology? A study of initial trust formation with organizational information systems. *The Journal of Strategic Information Systems* 17, 1 (2008), 39–71.

[26] Roger C Mayer and James H Davis. 1999. The effect of the performance appraisal system on trust for management: A field quasi-experiment. *Journal of applied psychology* 84, 1 (1999), 123.

[27] Roger C Mayer, James H Davis, and F David Schoorman. 1995. An integrative model of organizational trust. *Academy of management review* 20, 3 (1995), 709–734.

[28] Richard E Mayer and Valerie K Sims. 1994. For whom is a picture worth a thousand words? Extensions of a dual-coding theory of multimedia learning. *Journal of educational psychology* 86, 3 (1994), 389.

[29] D Harrison McKnight, Vivek Choudhury, and Charles Kacmar. 2002. Developing and validating trust measures for e-commerce: An integrative typology. *Information systems research* 13, 3 (2002), 334–359.

[30] D Harrison McKnight, Vivek Choudhury, and Charles Kacmar. 2002. The impact of initial consumer trust on intentions to transact with a web site: a trust building model. *The Journal of Strategic Information Systems* 11, 3 (2002), 297–323.

[31] D Harrison McKnight, Larry L Cummings, and Norman L Chervany. 1998. Initial trust formation in new organizational relationships. *Academy of Management review* 23, 3 (1998), 473–490.

[32] Bonnie M Muir. 1987. Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies* 27, 5-6 (1987), 527–539.

[33] Clifford Nass and Youngme Moon. 2000. Machines and mindlessness: Social responses to computers. *Journal of social issues* 56, 1 (2000), 81–103.

[34] Clifford Nass, Jonathan Steuer, and Ellen R Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems.* ACM, 72–78.

[35] Raja Parasuraman and Victor Riley. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human factors* 39, 2 (1997), 230–253.

[36] Darcy Podszebka, Candee Conklin, Mary Apple, and Amy Windus. 1998. Comparison of Video and Text Narrative Presentations on Comprehension and Vocabulary Acquisition. (1998).

[37] Byron Reeves and Clifford Nass. 1996. *How people treat computers, television, and new media like real people and places.* CSLI Publications and Cambridge university press.

[38] Julian B Rotter. 1967. A new scale for the measurement of interpersonal trust. *Journal of personality* 35, 4 (1967), 651–665.

[39] Julian B Rotter. 1980. Interpersonal trust, trustworthiness, and gullibility. *American psychologist* 35, 1 (1980), 1.

[40] Sim B Sitkin and Amy L Pablo. 1992. Reconceptualizing the determinants of risk behavior. *Academy of management review* 17, 1 (1992), 9–38.

[41] Nadaleen Tempelman-Kluit. 2006. Multimedia learning theories and online instruction. *College & Research Libraries* 67, 4 (2006), 364–369.

[42] Remo van der Heiden, Shamsi T Iqbal, and Christian P Janssen. 2017. Priming Drivers before Handover in Semi-Autonomous Cars. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems.* ACM, 392–404.

[43] Elke U Weber, Ann-Renee Blais, and Nancy E Betz. 2002. A domain-specific risk-attitude scale: Measuring risk perceptions and risk behaviors. *Journal of behavioral decision making* 15, 4 (2002), 263–290.

International Journal of Human–Computer Interaction

ISSN: 1044-7318 (Print) 1532-7590 (Online) Journal homepage: https://www.tandfonline.com/loi/hihc20

Investigating the Effect of System Reliability, Risk, and Role on Users' Emotions and Attitudes toward a Safety-Critical Drone System

Yusuf Albayram, Theodore Jensen, Mohammad Maifi Hasan Khan, Ross Buck & Emil Coman

To cite this article: Yusuf Albayram, Theodore Jensen, Mohammad Maifi Hasan Khan, Ross Buck & Emil Coman (2018): Investigating the Effect of System Reliability, Risk, and Role on Users' Emotions and Attitudes toward a Safety-Critical Drone System, International Journal of Human–Computer Interaction, DOI: 10.1080/10447318.2018.1491665

To link to this article: https://doi.org/10.1080/10447318.2018.1491665

INTERNATIONAL JOURNAL OF HUMAN–COMPUTER INTERACTION
https://doi.org/10.1080/10447318.2018.1491665

Investigating the Effect of System Reliability, Risk, and Role on Users' Emotions and Attitudes toward a Safety-Critical Drone System

Yusuf Albayram[a], Theodore Jensen[a], Mohammad Maifi Hasan Khan[a], Ross Buck [b], and Emil Coman[c]

[a]Department of Computer Science and Engineering, University of Connecticut, CT, Storrs, USA; [b]Department of Communication, University of Connecticut, CT, Storrs, USA; [c]Health Disparities Institute, University of Connecticut Health Center, CT, Hartford, USA

ABSTRACT

In safety-critical systems, it is essential to communicate relevant information to facilitate decision-making, promote trust, and improve performance without overloading users. To explore the effect of system performance information on rational and emotional processing by users, we performed a between-subject experiment in which participants were asked to imagine themselves as a drone operator or system administrator in a high-, medium-, or low-risk scenario. Then, based on their imagined scenario and role, participants rated the relevance of four aspects of system reliability to decision-making with the system, as well as the expected intensity of the GREAT emotions. Results indicate that system performance information affected participants' reasoning differently depending on risk level. Moreover, participants had different perspectives depending on their role in the system. Those in administrator roles indicated higher respect ratings for those with a similar role. These findings demonstrate that contextual risk and a user's role can influence emotions and attitudes toward safety-critical computer systems.

KEYWORDS

Human-computer Interaction; Emotions; Risk

1. Introduction

Highly sophisticated and complex automated systems are increasingly being used in various safety-critical contexts (e.g., transportation, aerospace, defense) (Redmill & Rajan, 1996). Unfortunately, these systems are often not perfectly reliable (Parnas, Van Schouwen, & Kwan, 1990), as a plethora of unpredictable events can cause them to malfunction or fail (e.g., hardware problems, software bugs, environmental factors) (Knight, 2002; Lutz & Mikulski, 2003). Thus, the human user is an integral aspect of safety-critical systems, and effective judgment is necessary in order to preserve the safety of people, property or the environment (Knight, 2002). Although communicating information regarding various aspects of computation (e.g., execution state, input data quality, communication delay) in a timely manner can promote trust, communicating too much information can increase cognitive demands and overload the user. Moreover, because of the risk in safety-critical scenarios, users' emotions are likely to influence their decision-making with the system. Therefore, identifying what information regarding system performance needs to be communicated, as well as the effects of system performance information on specific kinds of rational and emotional processing, is critical for enhancing the outcomes of safety-critical human-computer collaborations.

Toward that, this study investigates the effect of risk, role, and different information about the system's reliability on users' emotions and attitudes. In a two-way, 3 × 2 factorial experiment, participants were asked to imagine themselves in a high-, medium-, or low-risk scenario as the operator or administrator of a drone system (communication errors, slow response time, hardware failures, software updates) to various thoughts about the system. We also examined the role of emotions in safety-critical human-computer interaction (HCI) by asking participants to indicate the expected intensity of the GREAT emotions (gratitude, respect, elevation, appreciation, and trust).

We found that system performance information influenced participants' reasoning differently depending on risk level. Participants in the high-risk scenario were more concerned about "communication errors," "slow response time," and "hardware failures" than those in the medium- and low-risk scenarios. Due to the potential negative consequences of "communication errors" and "slow response time," information about these errors was deemed the most important by a majority of participants. Based on participants' comments, we observed two different views regarding the communication of system state information. Some focused on the impact of failures, while others considered the frequency of failures. Furthermore, participants' roles appeared to influence their attitudes about the system. While those in the administrator role seemed to not want to overwhelm operators with unnecessary information about software updates that were applied to fix bugs, those in the operator role seemed interested to know about the limitations of the system (e.g., what bugs the software update fixed, whether those issues were successfully fixed). Finally, we found that administrator participants gave significantly higher respect ratings for other administrators/designers than did those in the operator role.

CONTACT Yusuf Albayram yusuf.albayram@uconn.edu

We believe these findings have important implications for human interaction with safety-critical computer systems. By understanding the effects of risk and role on users' emotions and attitudes, designers and researchers may enhance the quality of communication about a system's reliability. The details of our study are presented in the following sections.

2. Related work

Today, there is a growing trend toward incorporating highly complex automation into modern systems (Redmill & Rajan, 1996). Automated systems can reduce user workload and improve safety (Satchell, 2016; Wiener, 1988), but they are sometimes not perfectly reliable. In safety-critical applications (e.g., drone operations, air traffic controllers), system reliability is salient because of the negative consequences associated with failure. Such accidents are well cited, prompting researchers to look for ways to improve automated systems used in safety-critical contexts (Knight, 2002; Parasuraman & Riley, 1997).

One way to address this problem is by designing more robust systems. Parnas (Parnas et al., 1990) highlights that in safety-critical systems, software can provide flexibility by allowing the system to handle many different inputs. However, because of the scale and complexity of the tasks, extensive testing, review, and documentation is needed throughout the design process to ensure that software is reliable. Similarly, Lutz and Mikulski (Lutz & Mikulski, 2003) note how insufficient design requirements can lead to system failure, and so evolution of software requirements following failures can make the system more robust in future situations. However, events not anticipated during the design process may still have irreversible consequences that can jeopardize the safety of people, property or the environment (Knight, 2002). As a result, human users are needed to take control of the system at times (Sheridan, 1992).

Moreover, there is evidence that increasing automation reliability does not necessarily improve performance. Prior work has demonstrated that high reliability automated systems may experience poor performance due to a lack of etiquette in their communication to users (Parasuraman & Miller, 2004) or due to "automation-induced complacency" (Molloy & Parasuraman, 1996; Singh, Tiwari, & Singh, 2009). Thus, designing for better reliability is not sufficient for the success of the system – the user's interests must also be considered.

Freedy et al. (Freedy, DeVisser, Weltman, & Coeyman, 2007) define this interaction between human and automated control as a *collaborative mixed initiative system*. Researchers on human-automation interaction also note that, compared to automation, humans are more flexible and "better equipped to respond to hanging or unpredictable conditions" (Parasuraman & Riley, 1997; Singh et al., 2009). The outcomes of such a collaboration thus depend on effective communication of task-relevant information between the involved parties. This includes communication of (1) system states and reliability to the user and (2) user thoughts and feelings to the system.

2.1. Information reduction vs. information overload

To improve system usability and reduce users' cognitive load, current system interfaces rely heavily on information reduction and hiding, rather than revealing too much system information to the user (Horvitz & Barry, 1995). Both *lack of transparency* and *information overload* can compromise a user's understanding of system capabilities and behavior, which can lead to *misuse* or *disuse*. *Misuse* entails reliance on an automated system that is insufficient for achieving some goal, while *disuse* represents a lack of reliance on a system that could actually help (Parasuraman & Riley, 1997).

Muir (Muir, 1994) was one of the first researchers to consider this as an issue of "trust" in an automated system. She noted that humans cannot ever have complete knowledge of a system's inner workings. However, a user must be able to predict certain system behaviors for a successful interaction (Muir, 1994). HCI researchers have subsequently explored this trade-off as a matter of users' "trust" toward machines (i.e., computers, robots, automation), applying human-human trust concepts to these automated trustees (Lee & See, 2004).

Prior work suggests that, to improve human-machine trust, information given to users should provide transparency, comprehensibility, and predictability of system actions (Beggiato et al., 2015; Bubb-Lewis & Scerbo, 1997; Choi & Ji, 2015; Itoh & Inagaki, 2004). Dzindolet et al. (Dzindolet, Peterson, Pomranky, Pierce, & Beck, 2003) investigated the notion of predictability, showing that observing system errors led study participants to distrust the system. However, giving explanations about when errors might occur increased trust and reliance. Furthermore, Antifakos et al. (Antifakos, Kern, Schiele, & Schwaninger, 2005) conducted a study where system confidence information was provided to study participants at various risk levels (e.g., low, medium, and high criticality). They found that displaying system confidence helped users to predict the system behavior, improving their trust in the system. In another study, Koo et al. demonstrated that explanations of "why" a system acts in a certain way (i.e., "Obstacle ahead") improved trust and driver attitudes, yet a "how" explanation (i.e., "The car is braking") was found to have a negative impact on trust (Koo et al., 2015). To the best of our knowledge, there is no prior work that has investigated the effect of risk and role on users' consideration of safety-critical system performance information.

2.2. User emotions

In situations involving risk and uncertainty, decision-making becomes more challenging not only because of the increase in cognitive workload, but also because of the strong effect of emotions (Kahneman & Tversky, 1979; Loewenstein, Weber, Hsee, & Welch, 2001). In their prominent review of trust in automation research, Lee and See concluded that "information that forms the basis of trust" grows out of human users' analytic, analogical, and affective processes, with the affective processes tending to have a dominant influence on the others (Lee & See, 2004). Yet, limited prior work has looked into the emotions associated with safety-critical human-automation interaction.

We expect emotions to be particularly relevant to user behavior with safety-critical computer systems.

Buck (Buck, 2014) has suggested that the GREAT emotions, Gratitude, Respect, Elevation, Appreciation, and Trust, are continually, albeit unconsciously, expressed and exchanged during the course of human-human interaction. Their function is to maintain the dignity and self-respect of each interaction partner. Buck notes that the more each interaction partner gives away the GREAT emotions, the more they tend to receive them from their partner in return.

Although the presence of the GREAT emotions in HCI has not been studied, we believe that humans apply them to non-human, automated interaction partners in a similar fashion. Nass et al. have studied social responses to computers, finding that despite denying so, users apply social norms to their automated interaction partners (Nass, Steuer, & Tauber, 1994). For example, in one lab experiment, participants who evaluated a computer's performance on a different computer gave more negative ratings than those who evaluated its performance on the same computer, although they denied treating the computer politely after the experiment. Nass et al. suggested that this demonstrates humans using politeness toward computer interaction partners (Reeves & Nass, 1996).

In the context of human-computer trust-building, Parasuraman and Miller (Parasuraman & Miller, 2004) have investigated the role of "etiquette" expressed by an automated system. They define "etiquette" as a "set of prescribed and proscribed behaviors that permits meaning and intent to be ascribed to actions." They conducted an experiment where participants interacted with a partially-automated flight simulator and compared a rude and intrusive communication style (i.e., lack of etiquette) to a more patient one. They found that good automation etiquette significantly improved system performance and user trust, regardless of system reliability. We build upon this work by considering the communication of system performance information as a form of "etiquette" and exploring its influence on user emotions.

Specifically, in this study, we investigate users' GREAT emotions and their thoughts about 4 aspects of a safety-critical drone system's reliability across 2 roles and 3 risk levels. We aim to leverage this information to improve both a system's expressive communication to users about its performance, as well as its ability to adapt to user emotions.

3. Methodology

3.1. Study design

The goal of this study is to better understand the effect of safety-critical system performance information on human users' thoughts and feelings, which influence decision-making and, thus, the outcomes of the human-computer collaboration. Specifically, this study seeks to answer the following questions:

- How does level of risk influence users' attitudes about different aspects of system performance?
- How does a human's role with respect to a safety-critical system and task influence attitudes about different aspects of system performance?
- What information regarding system performance is important and relevant to users to improve usability in safety-critical systems?
- Are the GREAT emotions relevant to safety-critical human-computer interaction?

To answer these questions, we designed six hypothetical scenarios involving safety-critical drone operations. Among multiple possible safety critical technologies (e.g., smart grid, self-drive car, assisted robots, drones), this study uses drones as an example as they are utilized for diverse applications (e.g., purely entertainment, border patrol, war). The experiment was a 3 (risk level: high/medium/low risk) × 2 (role: drone operator/system administrator), between-subject factorial design in which participants were randomly assigned to one of the six hypothetical scenarios.

The three "risk levels" used in the study are as follows:

- ***High Risk***: The drone is over a battlefield, and decisions involve identifying enemy targets who may be innocent civilians.
- ***Medium Risk***: The drone is over a border region, and decisions involve arresting suspected illegal immigrants who may be innocent citizens.
- ***Low Risk***: The drone is over the ocean, and decisions involve identifying whale pods or non-interesting seals for a company.

The two "roles" used in the study are as follows:

- ***System Administrator***: The task involves managing a drone that is used by someone else (e.g., operator), and making sure the system is working/operating properly.
- ***System Operator***: The task involves making decisions with and operating a drone that is overseen by an administrator.

3.2. Survey

We designed a survey consisting of multiple parts as follows.

First, participants were asked to answer demographic questions (e.g., age, gender, and level of education) and to report their level of computer proficiency. They were then shown a video providing brief information about what drones are and how they can be used for different purposes. A screenshot taken from this video is shown in Figure 1. Following the video, participants were asked if they understood what drones are, as well as whether or not they had prior experience with drones (for either fun or professional reasons).

Subsequently, participants were randomly assigned to one of the six scenarios and asked to provide a written explanation about how the system is operated, how reliable the system is, what their respective roles and tasks are in the given scenario, and the risks associated with decisions to be made. The descriptions of the scenarios were identical with the exception

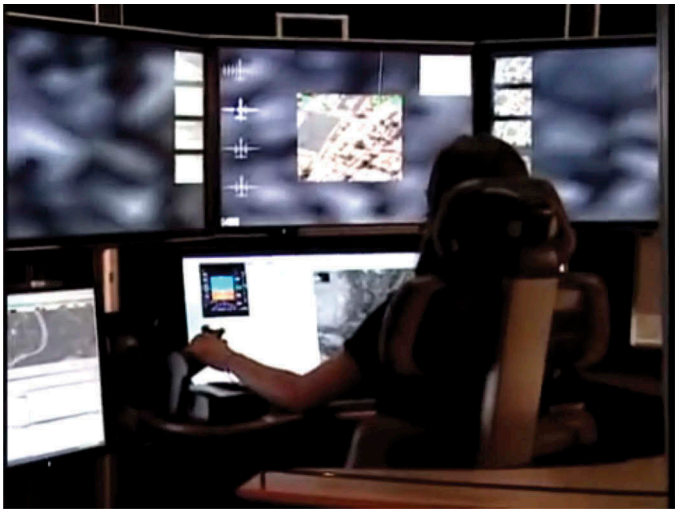

Figure 1. A screenshot from the video, showing how a drone system can be used in a safety-critical scenario (e.g., monitor battlefield). Please note that this screenshot was taken from a video that is publicly available on YouTube https://www.youtube.com/watch?v=unv9C2t7f5c. This survey has no association with the original video creators. We used only parts of the original video to provide an example of how a drone system can be used in a safety-critical scenario.

of the roles and risk situations they mentioned.[1] This was meant to ensure that participants understood the assigned scenario. The particular system was a drone system with some operational instabilities that could cause negative performance.

While considering the imagined scenario, participants were asked to rate the relevance of 4 different aspects of system reliability to various thoughts related to decision-making with the system. These four system state items indicate various aspects of the drone system's reliability:

- Recent software updates that were applied to fix unknown bugs
- Highly probable communication errors
- Occasional slow response time
- Low probability of hardware failures

These items were chosen because they communicate different types of errors an automated system may have, including failure to produce a response or message (i.e., communication errors), technological limitations leading to low accuracy (i.e., hardware failures, software bugs), or failure to respond at the right time (i.e., slow response time) (Singh et al., 2009).

We used five reasoning items adapted from the *CASC* (Communication Analytic and Syncretic Cognition) scale (Chaudhuri & Buck, 1997) to assess how system performance information influenced users' reasoning. The rational component of the *CASC* scale was comprised of the following five items:

- Make you think of X
- Make you think of pros and cons of X
- Make you think of arguments for or against regarding X
- Make you think of facts about X
- Make you think of facts about consequences of X

In our study, X is a statement reflecting the risk level of the participant's situation (e.g., innocent civilians are among the enemy target when using this system). For example, participants in the high-risk scenario were asked "*Would [slow response time] [make you think] whether [innocent civilians are among the enemy target], when using this system?*" on a scale ranging from 1 (definitely no) to 7 (definitely yes).

As these five items were found to be significantly positively correlated with each other, we combined them into a single, averaged "reasoning" item for each of the 4 pieces of system performance information. The index items were very reliable (Cronbach's α ranging from 0.902 to 0.918) (McKinley, Manku-Scott, Hastings, French, & Baker, 1997).

Moreover, to determine which system reliability information was most relevant to users, participants were asked to rank the importance of the system state information on a scale ranging from 1 (not at all important) to 7 (most important). Participants were also asked to provide open-ended commentary on their reasoning: "*Can you please explain the reason behind your answer to the above question in a few sentences?*"

Finally, to examine the relevance of the GREAT emotions to safety-critical human-computer interaction, participants were asked to rate the expected intensity of the GREAT emotions (gratitude, respect, elevation, appreciation, and trust) on a 7-point Likert scale.

3.3. Participants

We recruited participants from Amazon's Mechanical Turk (i.e., MTurk) platform, restricting participants to those 18 years of age or older, currently living in the United States, having greater than 1000 approved HIT's (Human Intelligence Tasks), and having HIT approval rate greater than 95%.

A total of 300 participants were recruited. However, we removed 4 responses from participants who failed to answer or did not provide a proper answer to the attention check question, which was asked to ensure that participants understood the scenario. Thus, a total of 296 valid responses were included in our analysis. Table 1 shows the distributions of participants in each risk level and role.

On average, participants took 17.7 minutes (*Median*= 14.8, *SD* = 11.6 minutes) to complete the survey for which they were compensated $3. The study was approved by the University's Institutional Review Board (IRB).

3.4. Demographics

Out of the 296 participants who completed the survey, 158 (53.4%) were male. Participants' age ranged from 19 to 67 with an average age of 33.5 years (median = 32, std = 9.4). All but 3 participants reported English as their native language.

Table 1. 6 hypothetical scenarios – 3 risk levels (i.e., high, medium and low risk) and 2 roles (i.e., administrator and operator). The number of participants in each scenario is also shown.

	Number of Participants	Role	Risk level
Scenario-1	48	System Admin	High
Scenario-2	51	System Operator	
Scenario-3	49	System Admin	Medium
Scenario-4	49	System Operator	
Scenario-5	50	System Admin	Low
Scenario-6	49	System Operator	

In terms of education level, 89.8% of participants reported having some form of postsecondary education (e.g., college or university), and the most frequently reported education level was a 4-year college degree 43.2% (128). The breakdown of the other reported education levels as follows: high school/GED (10.1%; 30), some college (23%; 68), 2 year college (14.9%; 44), master's degree (6.4%; 19), doctoral or professional degree (2.4%; 7). In terms of reported knowledge about computers in general, the majority of participants identified themselves as "proficient" 150 (50.7%) and "competent" 90 (30.4%), while 42 (14.2%) as "expert," 9 (3.0%) as "beginner," and 5 (1.7%) as "novice." Moreover, 7 (2.4%) participants reported that they did not know what drones are before watching the video. After watching the video, all but one participant reported that they understood what drones are. Overall, 39 (13.2%) participants reported having had experience with drones for either fun or professional reasons.

4. Findings

4.1. Sample statistics

To examine demographic differences among the six groups, we performed exploratory analysis with gender, age, level of education, knowledge about computers, and prior experience with drones. Results revealed that there were no significant differences in gender ($\chi^2(5) = 5.79$, $p = 0.32$), age ($\chi^2(5) = 4.93$, $p = 0.42$), education ($\chi^2(5) = 6.28$, $p = 0.27$), reported computer expertise ($\chi^2(5) = 7.86$, $p = 0.16$) nor having prior experience with drones ($\chi^2(5) = 5.12$, $p = 0.40$) between the groups.

Based on our analysis, we concluded that the six subgroups (scenario-1 to scenario-6) recruited were very similar in terms of demographics.

4.2. What information about system performance influences users' reasoning?

To examine the effect of system performance information on participants' reasoning across different scenarios and roles, we performed a two-way ANOVA with risk level (high, medium or low risk) and role (administrator or operator) for each averaged reasoning item corresponding to one of the 4 pieces of system performance information.

There was a significant main effect for "communication errors" with respect to risk level of the situation, $F(2,290) = 7.52$, $p < .001$. Participants in the high-risk scenario reported they would be more concerned about "communication errors" (*Mean* = 5.46, *SD* = 1.42) than participants in the medium-risk scenario (*Mean* = 5.42, *SD* = 1.35) and low-risk scenario (*Mean* = 4.76, *SD* = 1.44). A series of post hoc pairwise comparisons using Bonferroni correction revealed a significant difference in ratings between the high- and low-risk scenarios ($p = .002$), as well as those between the medium- and low-risk scenarios ($p = .004$). Moreover, as shown in Figure 2, participants in the system operator role generally gave higher ratings than those in the system administrator role. Mean ratings for system administrators were 5.33, 5.28, and 4.79, whereas the mean ratings for system operators were 5.59, 5.56, and 4.74 in high-, medium-, and low-risk situations, respectively.

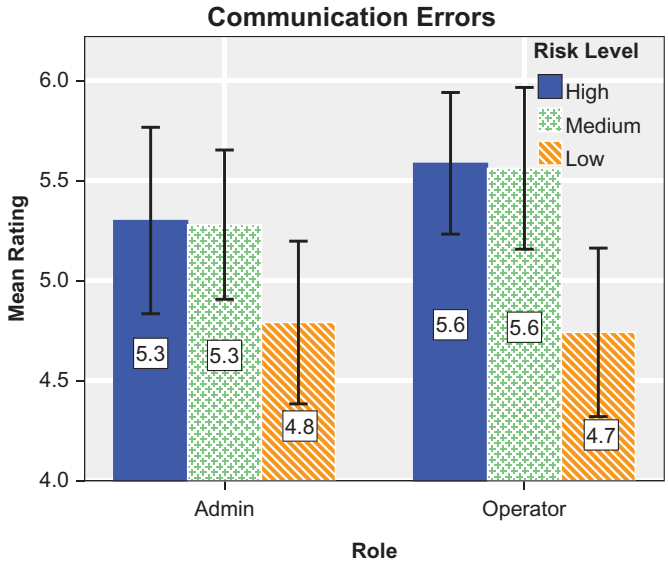

Figure 2. Mean ratings of the reasoning item for "communication errors" for the 3 risk levels and 2 roles. 95% confidence intervals are also included.

There was a significant main effect for "slow response time" with respect to risk level of the situation, $F(2,290) = 9.88$, $p < .001$. Participants in the high-risk scenario reported they would be more concerned about "slow response time" (*Mean* = 5.06, *SD* = 1.45) than participants in the medium-risk scenario (*Mean* = 4.40, *SD* = 1.57) and low-risk scenario (*Mean* = 4.16, *SD* = 1.38). A series of post hoc pairwise comparisons using Bonferroni correction revealed a significant difference in ratings between the high- and low-risk scenarios ($p < .001$), as well as between the high- and medium-risk scenarios ($p = .005$). Moreover, as shown in Figure 3, participants in the system operator role generally gave higher ratings than those in the system administrator role. Mean ratings for system administrators were 4.91, 4.35, and 4.27, whereas the mean ratings for system operators were

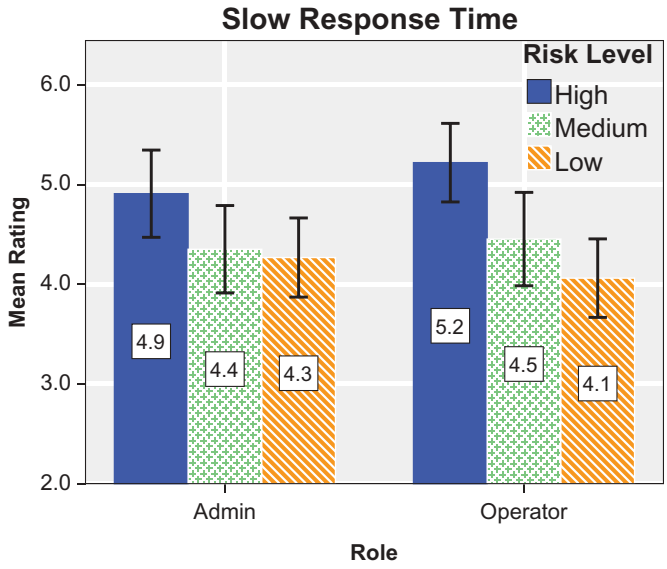

Figure 3. Mean ratings of the reasoning item for "slow response time" for the 3 risk levels and 2 roles. 95% confidence intervals are also included.

5.22, 4.45, and 4.06 in high-, medium-, and low-risk situations, respectively.

There was a significant main effect for "hardware failures" with respect to risk level of the situation, F (2,290) = 5.77, p = .003. Participants in the high-risk scenario reported they would be more concerned about "hardware failures" (*Mean* = 3.72, *SD* = 1.75) than those in the medium-risk scenario (*Mean* = 3.13, *SD* = 1.65) and low-risk scenario (*Mean* = 2.98, *SD* = 1.41). A series of post-hoc pairwise comparison using Bonferroni corrections revealed that a significant difference in ratings between the high- and low-risk scenarios (p = .004), as well as between the high- and medium-risk scenarios (p = .031). Moreover, as shown in Figure 4, participants in the system operator role generally gave higher ratings than those in the system administrator role. Mean ratings for system administrators were 3.49, 2.96, and 2.92, whereas the mean ratings for system operators were 3.95, 3.31, and 3.05 in high-, medium-, and low-risk situations, respectively.

There were no significant main effects for "software updates" with respect to risk level of the situation and role (see Figure 5). Also, there were neither main nor interaction effects for "role" for any of the system performance information.

These results suggest that a users' consideration of different system performance aspects depends heavily on risk level, such that for "communication errors," "slow response time," and "hardware failures," participants were more concerned about the system being used in higher-risk situations. However, level of risk did not influence the effect of information about "software updates" on participants' reasoning. This could be because the statements about "communication errors," "slow response time," and "hardware failures" emphasize possible failures of the system (e.g., highly probable communication errors, occasional slow response time and low probability of hardware failures), while the statement regarding "software updates" indicates that the software updates were already applied to fix unknown bugs. Thus, participants' level of concern about "software updates" did not vary significantly depending on the risk level of the situation.

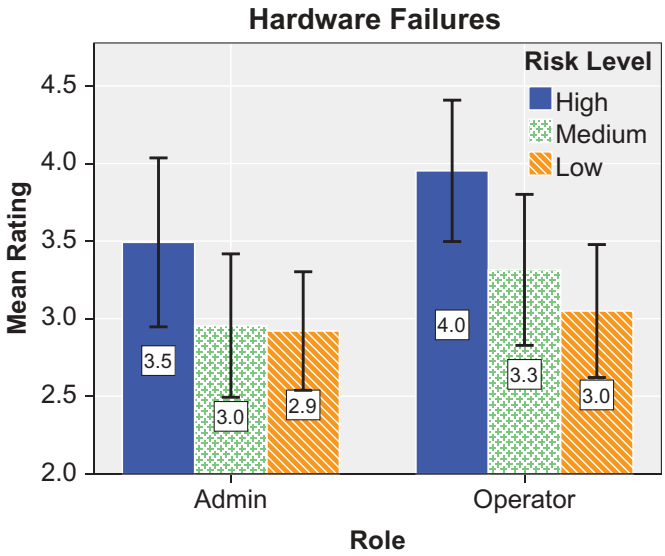

Figure 4. Mean ratings of the reasoning item for "hardware failures" for the 3 risk levels and 2 roles. 95% confidence intervals are also included.

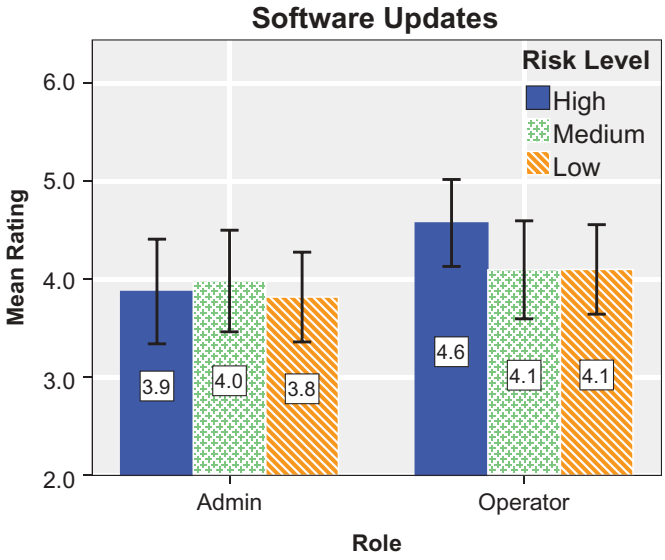

Figure 5. Mean ratings of the reasoning item for "software updates" for the 3 risk levels and 2 roles. 95% confidence intervals are also included.

4.3. What information regarding system performance is important?

To examine the importance of different system performance information to users in different scenarios and roles, we performed a two-way ANOVA with risk level (high, medium, or low risk) and role (system administrator vs. system operator) for each piece of system performance information.

Results show that "risk level" has a main effect for "slow response time" (F(2,290) = 5.64, p = .004) and "communication errors" (F(2,290) = 4.09, p = .018). Participants in the high-risk scenario (*Mean* = 5.72, *Med* = 6) ranked the expected severity of "slow response time" as more important than those in the medium-risk scenario (*Mean* = 5.11, *Med* = 5) and low-risk scenario (*Mean* = 5.21, *Med* = 5). Similarly, the expected severity of "communication errors" was rated as more important by participants in the high-risk scenario (*Mean* = 6.12, *Med* = 6) than participants in the medium-risk scenario (*Mean* = 6.09, *Med* = 6) and low-risk scenario (*Mean* = 5.68, *Med* = 6).

Moreover, "role" has a main effect for "software updates" (F(1,290) = 6.90, p = .009). Participants in the role of system operator (*Mean* = 5.18, *Med* = 5) ranked the information about recent software updates as more important than participants in administrator role (*Mean* = 4.71, *Med* = 5). This result shows two different perspectives based on a human's role with respect to a safety-critical system. In particular, participants in the administrator role may not want to overwhelm system operators with information about problems that were already fixed. Thus, from their point of view, the information about software updates was less important. On the other hand,

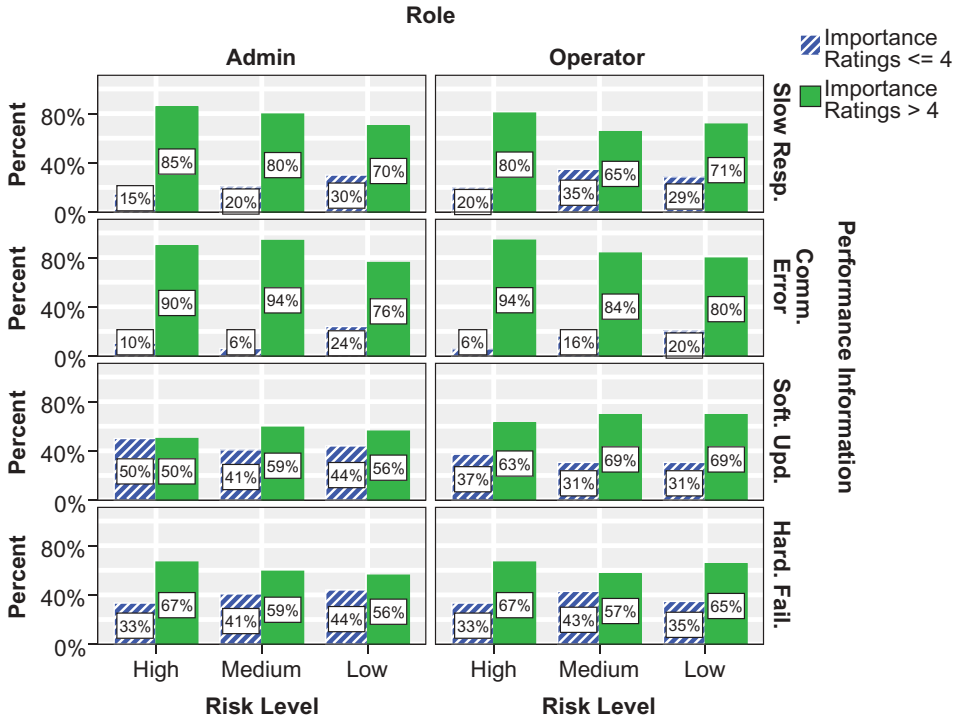

Figure 6. The percentage of participants rated the importance of corresponding information greater than 4 or lower than 5 on a scale ranging from 1 (not at all important) to 7 (most important).

participants in the system operator role seemed interested to know what bugs the software update fixed and whether or not those issues were successfully fixed. Thus, they would know the limitations of the system and be able to anticipate the likelihood of future problems (i.e., reliability and predictability).

We found neither main nor two-way interaction effects between "role" and "risk level" for the expected probability of "hardware failures." It is possible that participants were not concerned about hardware failures because they were informed that the probability of them occurring was low (e.g., once every 6 months). Figure 6 shows the percentage of participants who rated the importance of corresponding information greater than 4 or lower than 5 on a scale ranging from 1 to 7 (most important).

4.4. Correlation between reasoning and importance of system performance information

To understand how participants' reasoning related to the perceived importance of different aspects of system performance, we performed a correlation analysis between the reasoning item (i.e., the influence) and importance of system performance information for each of the four pieces of system performance information. Using Spearman's coefficients, we found that the reasoning items were significantly correlated with the importance of corresponding system performance information. Specifically, participants who rated the expected severity of "communication errors" as more important system performance information were likely to be more concerned about "communication error" ($\rho = 0.452$, $p < .001$). Similarly, participants who rated the expected severity of "slow response time" as more important were likely to be more concerned about "slow response time" ($\rho = 0.374$, $p < .001$). The same was true for "hardware failures" ($\rho = 0.216$, $p < .001$) and "software updates" ($\rho = 0.353$, $p < .001$).

4.5. GREAT emotions

To examine the relevance of the GREAT emotions to safety-critical human-computer interaction, we first analyzed the correlations among the GREAT emotions (i.e., Gratitude, Respect, Elevation, Appreciation, and Trust). Using Spearman's coefficients, we found that all the GREAT emotions were significantly correlated with each other. Specifically, participants who trust the system to function correctly are more likely to be grateful ($\rho = 0.650$, $p < .001$), appreciate ($\rho = 0.610$, $p < .001$), have respect for the others who manage the system ($\rho = 0.552$, $p < .001$) and less likely to be suspicious of the system ($\rho = -0.410$, $p < .001$), and vice versa (see Table 2).

Next, we performed a two-way ANOVA for each of the GREAT emotions to observe the effect of different risk levels and roles. We found a main effect for role ($F(1,289) = 4.89$, $p = .028$) with regard to ratings of respect. Participants in the system administrator role had higher respect ratings (*Mean* = 4.54, *Med* = 5) than those in the system operator role (*Mean* = 4.00, *Med* = 4). Participants in the system administrator role were asked to indicate their level of agreement with "I would have respect for the other system administrators/designers who work with you to manage the system," whereas participants in the system operator role rated their level of agreement with "I would have respect for the system administrators/designers who manage the system." This indicates that participants with similar administrator responsibilities

Table 2. Correlations among the GREAT emotions.

	Trust	Respect	Grateful	Appreciate
Respect	0.552**			
Grateful	0.650**	0.804**		
Appreciate	0.610**	0.818**	0.895**	
Suspicious	−0.410**	−0.179*	−0.257**	−0.252**

*Correlation is significant at the 0.01 level.
**Correlation is significant at the 0.001 level.

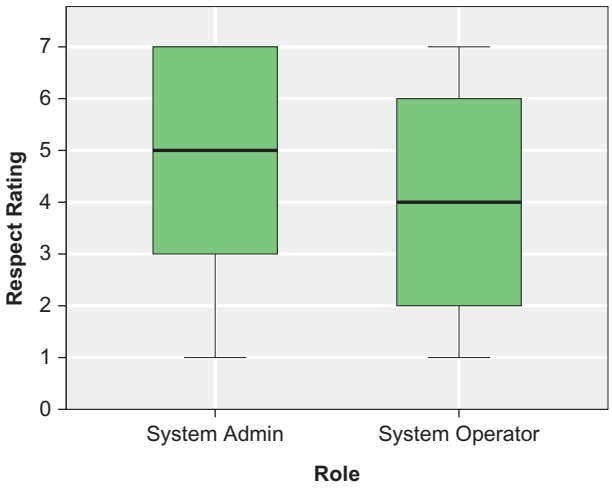

Figure 7. Box plot showing the median and interquartile range of rating of participants who were assigned to system administrator and system operator roles.

indicated higher respect ratings for the system administrators/designers than participants in the system operator role. Figure 7 shows the distributions (the median and interquartile range (IQR)) of ratings of participants assigned to the system administrator and system operator roles.

5. Discussion

5.1. Effect of system performance information on users' attitudes toward the system

Results show that information about system performance influenced participants' reasoning differently depending on risk level. Specifically, participants in the high-risk scenario were found to be significantly more concerned about "communication errors," "slow response time," and "hardware failures." Concern about "software updates" was not significantly different across risk levels. This could be because "communication errors," "slow response time," and "hardware failures" emphasize shortcomings of the system, whereas the statement regarding "software updates" indicates an improvement that was made to the system. Another possible explanation is that many participants tend to ignore "software updates" due to unclear benefits and past negative updating experiences (Fagan, Khan, & Buck, 2015; Vaniea & Rashidi, 2016). This was reflected in participants' comments such as:

"I don't care much about software updates because I probably wouldn't have a full understanding of what they were saying." (Low Risk, System Operator)

"Software updates might have resolved the known bugs but there is no telling yet if they might have negatively affected the system in other ways." (Low Risk, System Operator)

"Software update is usually glossed over by most people. As long as it works, then it works. Merely explain what it fixed." (High Risk, System Administrator)

This highlights the importance of transparency and feedback regarding system states for influencing users' attitudes toward computer systems, especially given their consideration of the safety-critical risks.

It is possible that some participants overrated the risk in our particular scenario. For example, some participants in the low-risk condition (i.e., identifying whale pods) may have considered the situation to be very risky, since failure would mean they did not accomplish their only goal and could result in "job loss." Nevertheless, even in this artificial scenario-based methodology, our results revealed considerably diverse ratings depending on the group to which participants were assigned. We believe that this work is a useful starting point for enhancing our understanding of how contextual risk and a user's role with respect to a task affect emotions and attitudes toward safety-critical computer systems.

We also acknowledge that the context of drone use may have influenced participants' attitudes toward the imagined scenario. For example, the mention of a "battlefield" in the high-risk scenario and "immigration" in the medium-risk scenario may be contentious topics for participants in the United States. Institutional trust, or a user's trust in the institution behind a computer system, can influence trust in the system itself (Lee & See, 2004). Thus, participants' interpretations of the risk level in their scenario may have been influenced not only by the consequences of poor performance, but by their own opinions regarding this kind of drone use.

5.2. Perceived importance of different system performance information

Concerns about "communication errors" and "slow response time"

We found that each reasoning item was significantly correlated with the corresponding importance of system performance information. This suggests that users' informational needs are driven by concerns about the context-specific consequences of system failure, a notion that is particularly relevant for designers of safety-critical systems used for high-risk tasks.

A majority of participants considered "communication errors" as the most important system performance information, followed by "slow response time." More specifically, 92% of participants in the high-risk scenario, 89% in the medium-risk scenario and 78% in the low-risk scenario ranked the importance of information regarding the expected severity of "communication errors" higher than 5 on a scale ranging from 1 (not at all important) to 7 (most important). For the importance of information regarding the expected severity of "slow response time," 83%, 72%, and 71% of participants in the high-, medium-, and low-risk scenarios, respectively, ranked higher than 5.

Participants stated the reasons why they regarded "communication errors" and "slow response time" as more important system state information as follows:

"The severity of the communication errors was my chief concern. My second concern was slow response time. If the drone operating system is lacking, pictures could come in a few seconds late and innocent civilians could be hurt." (High Risk, System Administrator)

"Communication errors is really vital because if there is a interruption in communication, the correct targets might not be able to be picked. Slow response time is important because any lag might make the drone unable to fulfill its purpose." (High Risk, System Administrator)

"Communication is important as well as response time in regards to drones. You as a person are not there. Therefore, you expect this drone to communicate the video and other things to you." (Medium Risk, System Operator)

"I think that the biggest problem for me, as taking the viewpoint from the person in the scenario, would be the slow response time. If I spot some movement in the ocean out of the corner of my eye on the screen, and I need to quickly look over, I need to do it ASAP. However, if there is a lag in the response time, then I could lose the animal completely. The same goes for the communication error, but on a much smaller scale." (Low Risk, System Operator)

As suggested by participants' open-ended responses, the potential negative consequences of communication errors or slow response time were more salient than for other pieces of system state information. It is possible that information about hardware failures and software updates was less specific compared to that about communication errors and slow response time. The broad range of outcomes associated with hardware failures and software updates may have been harder to imagine and, thus, our findings should be interpreted with caution.

Impact vs. frequency of "hardware failures"

We observed that participants had two different views when assessing the importance of system state information. Some focused on the impact of failures, while others considered the frequency of failures to be more relevant. For example, the following comments demonstrate these two different views about "hardware failures":

"I think the hardware failures happen less often so they would not be a high priority. The other two errors [Slow response time and Communication errors] happen quite often so it would be important to tell the operator about them." (Low Risk, System Administrator)

"I think recent updates to software to fix bugs is probably the most FREQUENT thing that would happen, so I ranked it highest because of chance of frequency. I ranked the probability of hardware failures lowest in most situations because it is a low probability of happening. I would trust that this system would operate well most of the time." (Low Risk, System Administrator)

"I feel that the probability of hardware failure is of supreme importance because without the hardware properly functioning, the drone itself will not be operable. All of the other system states would be irrelevant if the hardware was completely down." (Low Risk, System Administrator)

"It is not as important to know the probability of hardware failure, because even that is only a probability so it may or may not happen." (High Risk, System Operator)

"I simply tried to provide a value to each [System state information] based on what I perceived to be the level of urgency or importance." (High Risk, System Administrator)

Operators' vs. administrators' views on "software update" information

We observed that participants in the system operator role and those in the system administrator role had completely different perspectives on the system. Participants' position relative to the other agents in the system had an effect on their attitudes about different system performance aspects. Specifically, regardless of the risk level, operators and administrators had different views on the importance of "software update" information. Table 3 shows several sample comments of these participants.

These comments reflect the tradeoff between information reduction and overload in the face of a complex system. Prior work has pointed out that it is not uncommon for administrators to have incomplete mental models of the systems they manage (Barrett, Maglio, Kandogan, & Bailey, 2004), and administrators are often hesitant to apply software updates, which can potentially alter system behavior and the experience of end-users (Zhou et al., 2007). As administrators themselves are often not comfortable updating the system software, it may be the case that they feel like the operators will be overwhelmed by such information, explaining why they may be reluctant to convey such information.

While we found that risk level and role appear to influence participants' consideration of system performance factors, qualitative results shed light on the diverse perceptions held by individuals of such a system. Past experiences with technology are likely

Table 3. Sample comments of participants in the system administrator role and those in the system operator role on the importance of software update information.

	Administrator	Operator
High Risk	*"I would share information that would be relevant to the operation of the drone. No need to overwhelm them with information about things they will not effect them or things that are all ready fixed. I would rather they remember the things that will allow them to most effectively operate the drone."* (High Risk, System Administrator)	*"Foremost I would like to know what problems there were with the system that required software updates to fix and exactly what the update was purporting to fix."* (High Risk, System Operator)
Medium Risk	*"The information about software updates is less important from their [operators'] perspective. This is important to me as I have to maintain the systems, but typically, users are not interested in such specifics and usually they need not be."* (Medium Risk, System Administrator)	*"The knowledge that there have been updates or fixes to solves issues with bugs in the system is also quite important in that I can properly assess the potential weakness and deficiency of the systems in order to make adjustments as needed so as not to let that affect my reliance on the system."* (Medium Risk, System Operator)
Low Risk	*"The software update info doesn't have much to do with the day-to-day operations so I think that would be irrelevant to the operator."* (Low Risk, System Administrator)	*"I think that it's most important to keep the software updated so that any known problems are fixed. It is also important to expect problems to happen."* (Low Risk, System Operator)

to inform users' notions of different aspects of system performance as well. For example, participants' opinions about software updates were particularly apparent, as software updates often occur in personal computing settings. Similarly, the prevalence of mobile devices for communication may have informed opinions of communication errors and slow response time. We encourage future work to explore the role of various individual differences and experiences in shaping perceptions of safety-critical systems, as this has implications for the efficacy of transparency.

5.3. Presence of the GREAT emotions

The GREAT emotions (i.e., Gratitude, Respect, Elevation, Appreciation, and Trust) were found to be significantly correlated with each other, indicating the validity and reliability of these emotions. Initially, we expected that the GREAT emotions would be significantly different depending on risk level as well as role in the respective scenario. As the imagined drone system was imperfect and prone to err, participants were expected to express lower gratitude, respect, appreciation, and trust ratings along with higher elevation ratings in the high-risk scenario compared to the low-risk scenario.

We found that respect ratings significantly differed across roles, with administrator participants indicating higher respect ratings for other administrators/designers than participants in the operator role. This suggests that the system administrators took into account the responsibilities of a system administrator, such as being responsible for failure of the entire task, which led to higher respect ratings. It also suggests that human agents in a safety-critical system more positively regard those in roles similar to their own.

We found no significant main effects regarding risk level and role for the gratitude, appreciation, trust or elevation emotions. A likely cause of this is that the drone system in our study was imagined. Participants did not interact with a clear, computer entity beyond what they could envision in their mind. It is also important to acknowledge that the "respect" item specifically referred to system administrators and designers, while other GREAT emotions were stated relative to the computer system itself. These represent two distinct yet possibly related feelings. In our study, it is likely that emotions toward other human agents were more salient because descriptions of the roles gave specific mention of those other agents. Meanwhile, the lack of results for the other emotions is likely due to larger variability in participants' interpretations of the system, since it was more difficult to imagine given the lack of actual interaction. In addition, we acknowledge that because our study was based on self-reported ratings in response to a hypothetical human-computer interaction, the generalizability of our findings should be tested using real systems. However, we believe that this study has identified important aspects of communication strategies that can guide future research to improve the design of safety-critical systems.

Because humans treat computers as if they were social actors, even while acknowledging they are just machines (Nass et al., 1994), we believe that the GREAT emotions are likely to be expressed in HCI. Future work should explore the presence of the GREAT emotions in studies involving actual human interaction with a computer system, rather than the imagined scenario in our study.

Moreover, by addressing users' emotions as they interact with the system, the communication of relevant system information to users may be considered a form of etiquette, which has been found to influence trust and performance in safety-critical HCI (Parasuraman & Miller, 2004). Future work should further investigate how different kinds of expressive, emotionally-aware communication between a safety-critical system and its user may be leveraged to improve the outcomes of such collaborations.

6. Conclusion

Communicating information regarding safety-critical systems' reliability is one of the main factors that can influence the outcomes of human-computer collaborations. This study aimed to understand the effects of different information about the system's reliability on users' rational and emotional processing. Toward that, we executed a 3 × 2, between-subject factorial experiment in which participants were asked to imagine themselves in a high, medium, or low risk scenario in the role of a drone operator or system administrator. Participants rated the relevance of 4 different aspects of system reliability to various thoughts related to decision-making with the system, and how they would feel the GREAT emotions while imagining using the drone system.

Results indicate that participants' reasoning were affected differently depending on the risk level of the situation. Participants in the high-risk scenario were more concerned about "communication errors," "slow response time," and "hardware failures." Participants deemed "communication errors" and "slow response time" as the most important system performance information due to the impact of potential negative consequences. Moreover, we found that participant's role with respect to task had an influence which led to having completely different perspectives on the system. While participants in the administrator role seemed to not want to overwhelm system operators with unnecessary information (e.g., already fixed bugs), participants in the system operator role appeared to seek more information about the limitations of the system (e.g., what bugs fixed). In regard to users' emotions, administrator participants indicated significantly higher respect ratings than participants in the operator role due to greater responsibilities of a system administrator. We strongly believe that insight gained from this study will enable researchers to develop more effective expressive and scalable communication strategies for safety-critical systems.

Note

1. The entire written descriptions of scenarios are outlined in the Appendix.

Acknowledgment

This material is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-15-1-0490.

ORCID

Ross Buck http://orcid.org/0000-0001-7196-7794

References

Antifakos, S., Kern, N., Schiele, B., & Schwaninger, A. (2005). Towards improving trust in context-aware systems by displaying system confidence. In *Proceedings of the 7th international conference on human computer interaction with mobile devices & services* (9–14).

Barrett, R., Maglio, P. P., Kandogan, E., & Bailey, J. 2004. Usable autonomic computing systems: the administrator's perspective. In *Autonomic Computing, 2004. Proceedings. International Conference* (pp. 18–25).

Beggiato, M., Hartwich, F., Schleinitz, K., Krems, J., Othersen, I., & Petermann-Stock, I. (2015).What would drivers like to know during automated driving? Information needs at different levels of automation. In *Proceedings of the 7th conference on driver assistance*.

Bubb-Lewis, C., & Scerbo, M. (1997). Getting to know you: Human-computer communication in adaptive automation. In M. Mouloua & J. M. Koonce, *Human–Automation interaction: Research and practice* (pp. 92–99), Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Buck, R. (2014). *Emotion: A biosocial synthesis*. Cambridge, United Kingdom: Cambridge University Press.

Chaudhuri, A., & Buck, R. (1997). Communication, cognition and involvement: A theoretical framework for advertising. *Journal of Marketing Communications*, *3*(2), 111–125. doi:10.1080/135272697345998

Choi, J. K., & Ji, Y. G. (2015). Investigating the importance of trust on adopting an autonomous vehicle. *International Journal of Human-computer Interaction*, *31*(10), 692–702.

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. doi:10.1016/S1071-5819(03)00038-7

Fagan, M., Khan, M. M. H., & Buck, R. (2015). A study of users' experiences and beliefs about software update messages. *Computers in Human Behavior*, *51*, 504–519. doi:10.1016/j.chb.2015.04.075

Freedy, A., DeVisser, E., Weltman, G., & Coeyman, N. (2007). Measurement of trust in human-robot collaboration. In *Collaborative technologies and systems, 2007. cts 2007. international symposium on* (pp 106–114).

Horvitz, E., & Barry, M. (1995). Display of information for time-critical decision making. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence* (pp 296–305).

Itoh, M, & Inagaki, T. (2004). A microworld approach to identifying issues of human-automation systems design for supporting operator's situation awareness. *International Journal Of Human-computer Interaction*, *17*(1), 3–24.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society*, 263–291. doi:10.2307/1914185

Knight, J. C. (2002). Safety critical systems: Challenges and directions. In *Software engineering, 2002. icse 2002. proceedings of the 24rd international conference on* (pp 547–550). doi:10.1044/1059-0889(2002/er01)

Koo, J., Kwac, J., Ju, W., Steinert, M., Leifer, L., & Nass, C. (2015). Why did my car just do that? Explaining semi-autonomous driving actions to improve driver understanding, trust, and performance. *International Journal on Interactive Design and Manufacturing (Ijidem)*, *9*(4), 269–275. doi:10.1007/s12008-014-0227-2

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80. doi:10.1518/hfes.46.3.385.50404

Loewenstein, G. F., Weber, E. U., Hsee, C. K., & Welch, N. (2001). Risk as feelings. *Psychological Bulletin*, *127*(2), 267. doi:10.1037/0033-2909.127.2.267

Lutz, R. R., & Mikulski, I. C. (2003). Operational anomalies as a cause of safety-critical requirements evolution. *Journal of Systems and Software*, *65*(2), 155–161. doi:10.1016/S0164-1212(02)00057-2

McKinley, R. K., Manku-Scott, T., Hastings, A. M., French, D. P., & Baker, R. (1997). Reliability and validity of a new measure of patient satisfaction with out of hours primary medical care in the united kingdom: Development of a patient questionnaire. *Bmj*, *314*(7075), 193. doi:10.1136/bmj.314.7075.193

Molloy, R., & Parasuraman, R. (1996). Monitoring an automated system for a single failure: Vigilance and task complexity effects. *Human Factors*, *38*(2), 311–322. doi:10.1177/001872089606380211

Muir, B. M. (1994). Trust in automation: Part i. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics*, *37*(11), 1905–1922. doi:10.1080/00140139408963681

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. In *Proceedings of the sigchi conference on human factors in computing systems* (pp 72–78).doi: 10.3168/jds.S0022-0302(94)77044-2

Parasuraman, R., & Miller, C. A. (2004). Trust and etiquette in high-criticality automated systems. *Communications of the ACM*, *47*(4), 51–55. doi:10.1145/975817

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*(2), 230–253. doi:10.1518/001872097778543886

Parnas, D. L., Van Schouwen, A. J., & Kwan, S. P. (1990). Evaluation of safety-critical software. *Communications of the ACM*, *33*(6), 636–648. doi:10.1145/78973.78974

Redmill, F., & Rajan, J. (1996). *Human factors in safety-critical systems*. Newton, MA: Butterworth-Heinemann.

Reeves, B., & Nass, C. (1996). *How people treat computers, television, and new media like real people and places*. New York, NY: CSLI Publications and Cambridge university press.

Satchell, P. M. (2016). *Cockpit monitoring and alerting systems*. New York, NY: Routledge.

Sheridan, T. B. (1992). *Telerobotics, automation, and human supervisory control*. Cambridge, MA: MIT press.

Singh, A. L., Tiwari, T., & Singh, I. L. (2009). Effects of automation reliability and training on automation-induced complacency and perceived mental workload. *Journal of the Indian Academy of Applied Psychology*, *35*(2009), 9–22.

Vaniea, K., & Rashidi, Y. (2016). Tales of software updates: the process of updating software. In *Proceedings of the 2016 chi conference on human factors in computing systems* (pp 3215–3226).

Wiener, E. L. (1988). Cockpit automation. In Earl L. W. & David C. N (Eds.), *Human Factors in Aviation.* 433–461. San Diego, CA: Academic Press. ISBN: 978-0-08-057090-7

Zhou, Y., Marinov, D., Sanders, W., Zilles, C., d' Amorim, M., Lauterburg, S., Lefever, R. M., & Tucek, J. (2007). Delta execution for software reliability. In *Proceedings of the 3rd workshop on Hot Topics in System Dependability* 16. USENIX Association http://dl.acm.org/citation.cfm?id=1323140.1323156.

About the Authors

Yusuf Albayram is a post-doctoral fellow in the Computer Science and Engineering Department at the University of Connecticut where he received his PhD. His main research interests are in the interdisciplinary areas of usable security, human-computer interaction and ubiquitous computing.

Theodore Jensen is a PhD student in the Computer Science and Engineering Department at the University of Connecticut. His research interests include human-computer trust, affective computing, and the social impacts and dynamics of human-computer interaction.

Mohammad Maifi Hasan Khan is an Assistant Professor in the Department of Computer Science and Engineering at the University of Connecticut. He received his PhD degree in Computer Science from the University of Illinois, Urbana-Champaign. His research interests include usable security, risk communication, and performance modeling of large-scale systems.

Ross Buck is a Professor of Communication and Psychological Sciences at the University of Connecticut, Storrs. He is the author of Human Motivation and Emotion (Wiley, 1988), The Communication of Emotion (Guilford, 1984), and Emotion: A Biosocial Synthesis (Cambridge, 2014). His current interest is in the neuroscience of empathy.

Emil Coman is an emotional communication researcher with expertise in statistical modeling of complex causal processes. His research focuses on uncovering what causes health disparities, building, and interpreting evidence for healthcare providers' use, and translating complex statistical concepts for day to day use by researchers, students, and the general public.

Appendix

The Descriptions of the Scenarios:

Through a two-way, 3 (risk level: high risk/medium risk/low risk) × 2 (role: system operator/system administrator) factorial design experiment, participants were assigned one of the six hypothetical scenarios. Depending on the assigned role and risk level, participants were shown one of the phrasings separated by vertical bars (|) below. For instance, participants assigned to the system operator role in the high-risk scenario were shown ($Role_{opr}$) *and* ($Risk_{high}$), while participants assigned to the system administrator role in the low-risk scenario were shown ($Role_{adm}$) *and* ($Risk_{low}$), and so on. The entire written description of scenarios are outlined below.

Now, imagine that [$Role_{opr}$: *there are system administrators who are*] | [$Role_{adm}$: *you are the system administrator who is*] responsible for:

- Making sure that the software of the system that is used to operate the drone remotely is up-to-date.
- Making sure that the hardware of the system is up-to-date.
- Troubleshooting of the system if the performance is not acceptable.
- Performing preventative maintenance of the system.

However, despite [$Role_{opr}$: *their*] | [$Role_{adm}$: *your*] best effort, the system is not perfectly reliable and the system occasionally experiences the followings due to software bugs/hardware failures:

- The system occasionally crashes due to some unknown reasons and takes 2 minutes to reboot, making the system unavailable, and the timing/frequency of the crash is unpredictable.
- The system occasionally becomes very slow (e.g., freezes for 10 sec at a time) due to unknown software/hardware bugs.
- The system occasionally drops video frames due to communication errors.
- Different hardware components of the system rarely fails (e.g., once every 6 months).

Now, imagine that [$Role_{opr}$: *you are asked to use the system to make*] | [$Role_{adm}$: *the drone that you are responsible for managing is going to be used by someone else (e.g., operator) whose*] decisions involve identifying:

- $Risk_{high}$: enemy targets in a battlefield where there may be also innocent civilians.
- $Risk_{med}$: arresting or not arresting suspected illegal immigrants who may be innocent citizens in a border region.
- $Risk_{low}$: whale pods or non-interesting seals in the ocean for a company.

The Effects of Risk and Role on Users' Anticipated Emotions in Safety-Critical Systems

Yusuf Albayram[1(✉)], Mohammad Maifi Hasan Khan[1], Theodore Jensen[1], Ross Buck[2], and Emil Coman[3]

[1] Department of Computer Science and Engineering, University of Connecticut, Storrs, USA
{yusuf.albayram,maifi.khan,theodore.jensen}@uconn.edu
[2] Department of Communication, University of Connecticut, Storrs, USA
ross.buck@uconn.edu
[3] Health Disparities Institute, University of Connecticut Health Center, Hartford, USA
coman@uchc.edu

Abstract. Users of safety-critical systems often need to make risky decisions in real-time. However, current system designs do not sufficiently take users' emotions into account. This lack of consideration may negatively influence a user's decision-making and undermine the effectiveness of such a "human-computer collaboration." In a two-way, 2 (role: operator/system administrator) × 3 (risk level: high/medium/low) factorial study, we investigated the intensity of 44 emotions anticipated by 296 Mechanical Turk users who imagined being the (1) operator or (2) administrator of a drone system identifying (a) enemies on a battlefield, (b) illegal immigrants or (c) whale pods. Results indicated that risk level had a significant main effect on ratings of negative individualistic and negative prosocial emotions. Participants assigned to the high risk scenario anticipated more intense negative individualistic (e.g., nervous) and negative prosocial (e.g., resentful, lonely) emotions and less intense positive (e.g., happy, proud) emotions than participants assigned to the medium and low risk scenarios. We discuss the implications of our findings for the design of safety-critical systems.

Keywords: Emotions · Human-computer interaction
Decision-making · Risk

1 Introduction

Drone systems are increasingly being used for various purposes such as border patrol, battlefield monitoring, target tracking, and recreational activities. These systems can malfunction due to environmental factors, communication errors, or

D. Harris (Ed.): EPCE 2018, LNAI 10906, pp. 369–388, 2018.
https://doi.org/10.1007/978-3-319-91122-9_31

hardware and software failures, all of which may cause users to experience strong negative emotions (e.g., anger, anxiety, frustration, regret). Although there is a growing body of research showing that emotions strongly influence decision-making under risk and uncertainty [1–3], current safety-critical system designs do not consider users' emotions. This is likely to undermine effective decision-making, as strong emotions (e.g., regret, suspicion) can alter users' cognitive process [4].

While the role of emotion was long thought to be disruptive and contrary to models of decision-making, it is now understood that considering only the rational and cognitive is incomplete [5,6]. For instance, prior work in communication theory and psychology suggests that risky situations involve complex strong emotions (e.g., fear, suspicion, excitement) and that, if forewarned about what emotions to expect (i.e., emotional education), people are less surprised by their emotions [1,4,7]. This can allow for mindful processing of risks (e.g., emotional inoculation) [6]. Because of the risks faced by safety-critical system users, we argue that "emotional inoculation" is widely applicable to safety-critical human-computer interaction, and should be explicitly considered while designing user interfaces. A system that communicates about emotions can improve decision-making by allowing users to process the strong negative and positive emotions that arise in their safety-critical tasks. Before designing such systems, it is important to identify the relevant emotions.

As a first step towards this goal, we investigated the effect of risk level and role on users' anticipated emotions in a two-way, 2 (role: operator/system administrator) × 3 (risk level: high/medium/low) factorial experiment. We recruited 296 participants on Amazon's Mechanical Turk platform and provided them with a written description of one of six hypothetical scenarios where they were asked to imagine themselves as a drone operator or system administrator in a high, medium, or low risk scenario. Participants rated the anticipated intensity of 44 emotions in their scenario. Our findings show that risk level had a significant main effect on negative individualistic emotions and negative prosocial emotions. Participants in the high risk scenario expected more negative individualistic (e.g., nervous), more negative prosocial (e.g., resentful, lonely) and fewer positive (e.g., happy) emotions than participants in the medium and low risk scenarios. Insights gained in this study can enhance our understanding of the emotional aspects of decision-making in safety-critical human-computer interaction. The details of our study are presented in the following sections.

2 Background

2.1 Emotions in Decision-Making

Decision-making is the process of selecting a preferred option or course of action among a number of choices [8]. For a considerable time, decision-making was regarded by researchers as a predominantly cognitive process. According to utility theory, decision-makers evaluate the potential consequences of their options and choose the one they believe will yield the most beneficial result (i.e., the

"utility-maximizing" alternative) [9]. Research on decision-making in the last couple of decades has shown that this view is incomplete. There is now a significant amount of psychological research demonstrating that emotions influence decision-making in various ways [4,7].

In a review of these works, Loewenstein and Lerner [4] note two different ways in which emotions enter into a decision: (1) *expected emotions* and (2) *immediate emotions*. *Expected emotions* are those that a decision-maker thinks they will experience as a consequence of some decision. Considered alongside the utility model, the decision-maker will evaluate the consequences of their options and choose that which they expect to maximize positive emotions and minimize negative emotions. *Immediate emotions* are those experienced at the time of decision-making.

Prior work suggests that *immediate emotions* and *expected emotions* are interconnected: *immediate emotions* can impact expectations about future emotions, while *expected emotions* that are anticipated by a decision-maker can influence their current emotional state [4]. For instance, studies have shown that if a decision-maker is presently experiencing positive emotions, his or her evaluation of certain options is likely to be more positive, while those experiencing negative emotions are likely to make more negative evaluations [10,11]. This is exemplified by a "hot/cold empathy gap," in which individuals in a "hot" emotional state (e.g., angry) have been observed to poorly predict their feelings or behavior when in a "cold" state (e.g., not angry) [12]. Additionally, findings that positive emotions broaden attentional focus while negative emotions narrow it [13,14] suggest that the valence and nature of an individual's *immediate emotions* influence their cognitive processing. These dynamics have clear implications for decision-making.

In situations involving risk and uncertainty, not only is there a potential increase in cognitive workload, but the effects of the decision-maker's emotions become more pronounced [1,3]. The "*risk as feelings hypothesis*" explores this notion to explain behavioral responses that differ from what individuals cognitively view as the best course of action. While moderately intense emotions tend to play an "advisory role," and their influence on an individual's judgment can often be limited [4,15], strong emotions generally exert more control over behavior. The "*risk as feelings hypothesis*" lends this to the role of "anticipatory" emotions such as fear, worry, and anxiety as inputs in the decision-making process. Specifically, there are a different set of determinants for cognitive evaluations of risk and emotional reactions to risks. While the former is influenced by factors such as outcome probability and severity, emotions are influenced more so by the vividness of imagined consequences or experience with certain outcomes. For instance, feelings about risk have been found to be insensitive to changes in probability, contrary to cognitive evaluations of risk [1].

Use of safety-critical systems is a high-risk, decision-making context where both moderate, advisory emotions and stronger emotions are likely to be at play.

2.2 Emotions in Human-Computer Interaction

Safety-critical system users such as drone operators and air traffic controllers often need to make decisions under uncertainty and time pressure. As wrong decisions may lead to serious consequences for people, property and the environment [16], users of such systems are likely to experience strong anticipatory emotions. Likewise, although the probability of the computer system failing is likely to be low, the potential negative consequences can be emotionally salient. Therefore, it is important to understand what specific emotions may be experienced by users.

Interaction with computers is often portrayed as a purely cognitive endeavor, given that the machines literally operate based on logic. However, recent research highlights the importance of emotional considerations in human computer interaction, wherein a computer that can recognize human emotion can appropriately respond its user's emotions, thus improving the user experience and outcomes of the interaction [17–20]. In one application, Jones and Jonsson [21] proposed an emotionally responsive car system that tracks the emotional state of a driver based on their speech. This information is then used to modify the car's navigational voice, which can relax a tense driver or make them happier about the current conditions. This can improve the driver's concentration and improve safety. This study reports promising results on the potential for emotions to be actively and effectively leveraged in safety-critical human-computer interaction.

Recently, Buck et al. [22] presented the User Affective eXperience (UAX) scale, measuring self-reported emotions that were anticipated in response to pop-up software update messages. They reported 4 latent factors (positive affect, anxiety, hostility and loneliness) which were found to be significantly different between a pressured condition (imagining working on an urgent and stressful task) and a relaxed condition (imagining surfing on the Web while relaxing). Their findings suggest that considering only emotional valence is inadequate, while distinguishing between individualist and pro-social emotions can paint a more thorough picture of the dynamics of affect-influenced decision-making in HCI.

It is fairly obvious that the stress associated with risky, safety-critical system use may cause a user to experience individualistic emotions such as anger or confusion. It is less clear for prosocial emotions, such as guilt and shame, which are those associated with adherence to social norms and group cooperation [23]. First, these are relevant in the drone context because of the presence of other people: system use can have direct consequences for people on the ground, while human operators and administrators work together on tasks with the system. Yet further, a substantial amount of research showing that humans respond socially to computer interaction partners [24,25] suggests that prosocial emotions may arise in the "group cooperation" between human members of the team and the computer system itself. Whereas Freedy et al. [26] sought to define better performance metrics for the unique "interaction of two cognitive systems" (i.e., the human and the computer), we argue that human emotions play an equally important role in the dynamics of such a "collaborative mixed initiative system".

For example, user emotions may contribute to their "trust" in an automated system, which has been found to influence reliance decisions [27]. Problems of automation *disuse*, in which operators do not use a system when it may help, and *misuse*, in which operators use a system when it is insufficient for some task, are well cited and have been linked to poor "calibration" of trust by the user ("undertrust" and "overtrust," respectively) [28]. Thus, several researchers have investigated the factors that influence a trust in automation, often varying system reliability and measuring trust with self-reports [29]. While it has been noted that there may be affective components of trust in addition to analogical ones, the role of emotions in trust decisions has not been sufficiently studied. Given that the consequences to poor trust calibration may be particularly severe with safety-critical systems, we argue that affective trust is highly influential on users' decision-making.

While some research efforts have investigated the influence of emotions in human-computer interaction (HCI), to the best of our knowledge, we are the first to investigate the effects of risk and role on users' anticipated emotions in the context of safety-critical drone applications. Specifically, this study expands upon Buck et al.'s work [22] and explores the anticipated intensity of 44 discrete emotions across various roles and risk levels with respect to a safety-critical drone system.

3 Methodology

3.1 Study Design

This study investigates how a safety-critical system users' anticipated emotions vary depending on their role and the criticality of the situation. Toward that, we designed six hypothetical scenarios involving drone operations. Among multiple possible safety critical technologies (e.g., smart grid, self-driving car, assisted robots, drones), this study uses drone because they are utilized for diverse applications (e.g., purely entertainment, border patrol, war).

The experiment was a 2 (role: operator/system administrator) × 3 (risk level: high/medium/low), between-subject factorial design where participants were randomly assigned to one of six hypothetical scenarios. Participants were asked to rate the anticipated intensity of 44 emotions while imagining themselves in their "risk level" and "role."

The two "roles" used in the study are as follows:

- ***System Administrator*:** The task involves managing a drone that is used by someone else (e.g., operator), and making sure the system is working/operating properly.
- ***System Operator*:** The task involves making decisions with and operating a drone that is overseen by system administrators.

The three "risk levels" used in the study are as follows:

- ***High Risk*:** The drone was over a battlefield, and the decisions involve identifying enemy targets who may be innocent civilians.
- ***Medium Risk*:** The drone was over a border region, and the decisions involve arresting suspected illegal immigrants who may be innocent citizens.
- ***Low Risk*:** The drone was over the ocean, and the decisions involve identifying whale pods or non-interesting seals for a company.

The written descriptions of the scenarios were identical with the exception of the roles and risk level they mentioned, and are outlined in the Appendix. In particular, the hypothetical drone system had some operational instabilities that could cause negative performance. This information was intended to stimulate participants' emotional responses as they imagined making decisions in a safety-critical situation (i.e., with potentially dangerous consequences) with this imperfect system.

3.2 Survey

We designed a survey consisting of multiple parts as follows.

First, participants were asked to answer demographic questions (e.g., age, gender, and level of education) and report their level of computer proficiency. They were then shown a video about drones and their various applications. Following the video, participants were asked if they understood what drones are, and whether they had prior experience with drones (for either fun or professional reasons).

Subsequently, participants were randomly assigned to one of the six scenarios and, as an attention check, were asked to provide a written explanation of how the drone system is operated, how reliable it is, what their role and task was in the given scenario, and the risks associated with decisions they would have to make.

Finally, participants were asked to rate the expected intensity of 44 different emotions on a scale ranging from 1 (the least amount of intensity) to 7 (the greatest amount of intensity). The emotions were presented in the format "*I would feel* [***Emotion***]" and shown to participants in random order to avoid biasing them. These emotions were chosen to cover the broad range of emotional responses one could have while using a computer system [22,30,31]. The list of the 44 emotions can be seen in Table 3 in the Appendix.

We expected participants in the high risk scenario (i.e., identifying enemies on a battlefield) to report higher levels of negative emotions (e.g., nervous, anxious) than those in the medium risk (i.e., identifying illegal immigrants) and low risk (i.e., identifying whale pods) scenarios. Additionally, we expected the intensity of negative and positive emotions to vary between operator and administrator roles in the same scenario due to different responsibilities.

Moreover, prior work has found distinction between individualistic and prosocial emotions in response to pop-up software update warning messages [22]. In

our hypothetical context, the distinction between individualistic and prosocial emotions may also be salient, given that (1) system failure could lead to negative consequences for other people and (2) the task involves collaboration with other people and the computer system itself. Thus, we expected to find differences in individualistic and prosocial emotions across risk levels and roles.

3.3 Participants

We recruited participants from Amazon's Mechanical Turk (MTurk) platform. We restricted participants to those 18 or older, currently living in the United States, having greater than 1000 approved HIT's (Human Intelligence Tasks), and having a HIT approval rate greater than 95%.

A total of 300 participants were recruited. We removed the responses of 4 participants who failed to properly answer the attention check question. Thus, a total of 296 valid responses were included in our analysis. Table 1 shows the distributions of participants among the six groups.

Table 1. 6 hypothetical scenarios: 2 roles (i.e., administrator and operator) and 3 risk levels (i.e., high, medium and low risk). The number of participants in each group is also shown.

	Number of participants	Role	Risk level
Scenario-1	48	System admin	High risk
Scenario-2	51	System operator	
Scenario-3	49	System admin	Medium risk
Scenario-4	49	System operator	
Scenario-5	50	System admin	Low risk
Scenario-6	49	System operator	

Participants took an average of 17.7 min ($Median$=14.8, SD=11.6 min) to complete the survey and were compensated with $3. The study was approved by the University's Institutional Review Board (IRB).

3.4 Demographics

Out of 296 participants who completed the survey, 158 (53.4%) were male. Participants' age ranged from 19 to 67 with an average of 33.5 years (median = 32, std = 9.4). All but 3 participants reported English as their native language.

In terms of education level, 89.8% of participants reported having some form of postsecondary education (e.g., college or university) while the most frequent reported education level was a 4-year college degree 43.2% (128). The breakdown of the other reported education levels is as follows: high school/GED (10.1%; 30), some college (23%; 68), 2 year college (14.9%; 44), master's degree (6.4%; 19), and doctoral or professional degree (2.4%; 7).

In terms of reported knowledge about computers in general, 9 (3.0%) participants identified themselves as "beginner," 5 (1.7%) as "novice," 90 (30.4%) as "competent," 150 (50.7%) as "proficient," and 42 (14.2%) as "expert." Moreover, 7 (2.4%) participants reported that they did not know what drones were before watching the video, while only one participant reported not knowing after watching the video. Overall, 39 (13.2%) participants reported having had experience with drones for either fun or professional reasons.

To examine demographic differences among the six groups, we performed an exploratory analysis with gender, age, level of education, knowledge about computers, and prior experience with drones. The results of the analysis revealed no significant differences in gender ($\chi^2(5) = 5.79$, $p = 0.32$), age ($\chi^2(5) = 4.93$, $p = 0.42$), education ($\chi^2(5) = 6.28$, $p = 0.27$), reported computer expertise ($\chi^2(5) = 7.86$, $p = 0.16$) or prior experience with drones ($\chi^2(5) = 5.12$, $p = 0.40$) across the six groups.

Based on our analysis, we concluded that the groups recruited were similar in terms of demographics.

4 Findings

We first performed an exploratory Principal Component Analysis (PCA) on the ratings of the 44 anticipated emotions. This analysis allowed us to cluster the emotions into groups (i.e., factors) and determine the characteristics of each. Subsequently, for each factor extracted, we performed a 2-way, 2×3 (role $\times$ risk level) Analysis of Variance (ANOVA). The details are presented below.

4.1 Factor Analysis

To assess the appropriateness of the collected emotion data for factor analysis, we first conducted several diagnostic tests using well-known sampling adequacy measures. Bartlett's test of sphericity measure is ($\chi^2(946) = 8725.2$, $p < 0.0001$) and the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy is 0.934. According to the Kaiser criterion, 0.9 and above reveals marvelous value [32], suggesting that our data was correlated and that the variability can be explained by common factors.

Subsequently, we conducted an exploratory PCA on the ratings of the 44 emotions and extracted 6 emotion factors based on the Kaiser criterion (i.e., K1 rule: retain factors if eigenvalue is greater than 1). However, as the Kaiser criterion often leads to substantial overfactoring [33], we also performed parallel analysis and determined the optimal coordinates. Briefly, parallel analysis calculates eigenvalues based on the same sample size and number of variables using sets of random data. Then, each *ith* eigenvalue obtained from the random data is compared with the *ith* eigenvalue produced by the actual data. Based on this comparison, the eigenvalue is retained if the eigenvalue expected from random data is greater than the eigenvalue calculated by the factor analysis. The optimal coordinate method uses linear regression to determine the coordinates where an

eigenvalue diverges [34]. These two methods (i.e., parallel analysis and optimal coordinates) are widely used for determining the appropriate number of factors.

As shown in Fig. 1, both parallel analysis and optimal coordinates suggest extracting three factors for our data. Based on the aforementioned methods, we extracted three factors. These three factors predicted a cumulative total of 55.78% of the variance where factors 1, 2 and 3 explain 29.51%, 18.97%, and 7.29% of the variance, respectively.

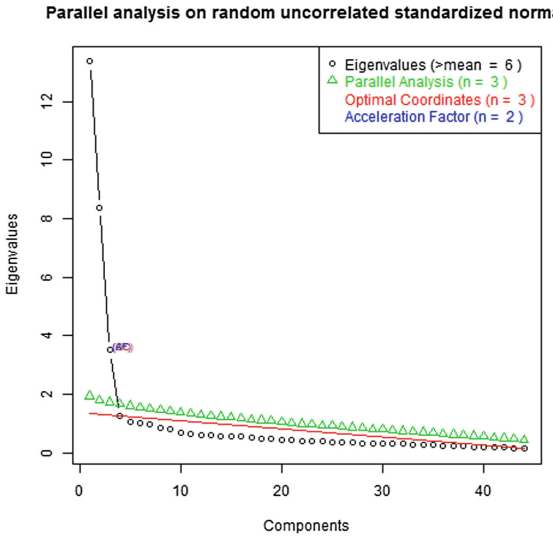

Fig. 1. Scree plot showing eigenvalues from the factor analysis, parallel analysis, optimal coordinates, and acceleration factor.

We used Varimax (orthogonal) as the rotation method, wherein prior work has considered items with a loading above 0.4 to be loaded on a factor [35]. Table 2 shows the rotated factor loadings of 44 emotions as well as the emotions belonging to each factor. Nineteen emotions such as angry, nervous and dismayed were included in Factor-1, which was labeled as "Negative individualistic" emotions. Fifteen emotions such as happy, welcomed and grateful were included in Factor-2, which was labeled as "Positive" emotions. Lastly, ten emotions such as scornful, disdainful and resentful were included in Factor-3, which was labeled as "Negative prosocial" emotions. These factors support those found in Buck et al.'s work [22] in the context of software update pop-up warnings, with our "Negative individualistic" corresponding to their "Anxious," our "Positive" to that of the same label, and "Negative prosocial" to the pair of factors "Lonely" and "Hostile."

For our three extracted factors, we also computed reliability measures using Cronbach's α. As shown in the second to last row of Table 2, all Cronbach's α values are higher than 0.7. According to McKinley et al. [36], $\alpha > 0.6$ indicates

Table 2. Factor loadings of the 44 emotions from the factor analysis. The highest factor loadings of each factor are highlighted in bold to facilitate visualization. The reliability measures (Cronbach's α and average inter-item correlation (IIC)) are also shown in the last two rows.

	Factor-1	Factor-2	Factor-3
Angry	**.775**		
Nervous	**.773**		
Dismayed	**.771**		
Anxious	**.768**		
Distraught	**.751**		
Ashamed	**.746**		
Down	**.732**		
Embarrassed	**.731**		
Afraid	**.731**		
Guilty	**.716**		
Sad	**.697**		
Freaked out	**.664**		
Depressed	**.649**		
Disgusted	**.632**		
Confused	**.592**		
Dazed	**.587**		
Hostile	**.528**	.475	
Isolated	**.514**	.484	
Surprised	**.444**		
Happy		**.776**	
Welcomed		**.769**	
Grateful		**.762**	
Admiring		**.760**	
Proud		**.758**	
Triumphant		**.758**	
Powerful		**.756**	
Secure		**.746**	
Trusting		**.744**	
Friendly		**.737**	
Cared-for		**.730**	
Respectful		**.717**	
Confident		**.681**	
Vigorous		**.672**	
Energetic		**.668**	
Scornful			**.789**
Disdainful			**.745**
Resentful			**.729**
Dishonored			**.716**
Contemptuous			**.709**
Humiliated			**.669**
Arrogant			**.651**
Lonely			**.602**
Insulted	.493		**.571**
Abandoned	.504		**.515**
Cronbach's alpha (α)	.946	.940	.886
IIC	.479	.511	.525

satisfactory internal reliability for all sub-scales. Finally, we calculated average inter-item correlation (IIC) values. As shown in the last row of Table 2, all the sub-scales are above 0.30, indicating "exemplary" reliability [37]. Based on our analysis, we concluded that each of the extracted factors had high reliability.

4.2 ANOVA Analysis

As we wanted to better understand how users might feel while using the drone system in different scenarios and roles, we performed a two-way, (2 × 3) ANOVA for each emotion factor extracted from the factor analysis. More specifically, the dependent variables for our three ANOVAs were negative individualistic emotions (factor-1), positive emotions (factor-2), and negative prosocial emotions (factor-3). We included risk level (high, medium, and low risk), role (system operator and system administrator), and their interaction effects as independent variables in each analysis. The details are presented below.

The ANOVA revealed that risk level had a significant main effect on negative individualistic emotions $F(2,290) = 6.8$, $p = .001$ and negative prosocial emotions $F(2,290) = 4.1$, $p = .017$. Participants assigned to the high risk scenario anticipated stronger negative individualistic emotions (e.g., nervous, confused) and negative prosocial emotions (e.g., resentful, lonely), but weaker positive (e.g., happy, grateful) emotions than those assigned to the medium risk and low risk scenarios. More specifically, participants in the high risk scenario ($Mean = 3.54$, $SD = 1.27$) rated higher negative individualistic emotions than participants in the medium scenario ($Mean = 3.11$, $SD = 1.32$) and the low risk scenario ($Mean = 2.88$, $SD = 1.26$). A series of post-hoc pairwise comparisons using Bonferroni correction revealed that the difference in ratings between the

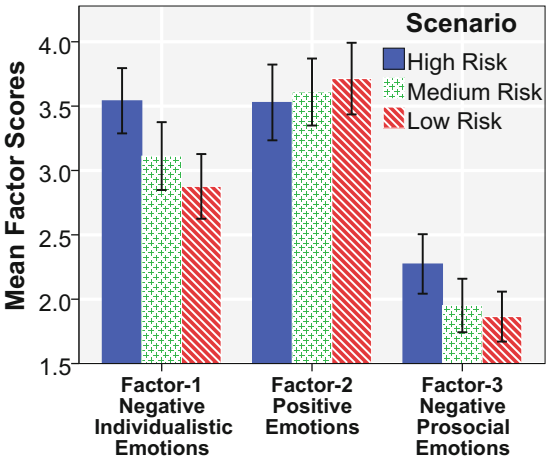

Fig. 2. Mean factor scores for the three risk levels (high risk/medium risk/low risk) for each factor. 95% confidence intervals are also included.

high risk and the low risk scenarios was significant ($p < .001$). Similarly, participants in the high risk scenario ($Mean = 2.27$, $SD = 1.16$) rated negative prosocial emotions higher than participants in the medium risk scenario ($Mean = 1.95$, $SD = 1.04$) and the low risk scenario ($Mean = 1.86$, $SD = 0.97$). A series of post-hoc pairwise comparisons using Bonferroni correction revealed that the difference in ratings between the high risk and low risk scenarios was significant ($p < .021$). Although those in the high risk scenario ($Mean = 3.53$, $SD = 1.48$) rated lower levels of positive emotions than participants in the medium risk scenario ($Mean = 3.61$, $SD = 1.30$) and low risk scenario ($Mean = 3.71$, $SD = 1.40$), the difference in ratings among the three risk levels was not statistically significant. The mean factor scores for the three risk levels are shown in Fig. 2.

The ANOVA also revealed that there was no significant main effect on emotions due to role. The mean factor scores for the two roles (operator/system administrator) can be seen in Fig. 3.

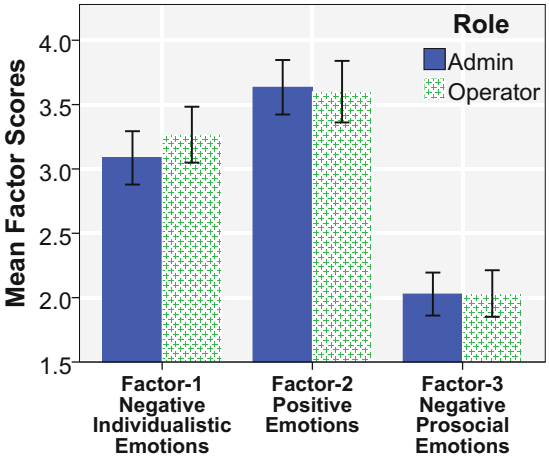

Fig. 3. Mean factor scores for the two roles (system administrator/system operator) for each factor. 95% confidence intervals are also included.

5 Discussion

Despite a growing body of literature demonstrating the significant role of emotions in the decision-making process, we have a relatively limited understanding of the specific emotions relevant to high risk decision-making. As safety-critical technologies such as drones and self-driving cars become more prevalent, so will the high-risk decisions to which their users must attend. To gain insight into the effect of risk level and role on safety-critical system users' emotions, we asked participants to imagine themselves as a drone operator or system administrator

in a high, medium, or low risk scenario. They then rated the expected intensity of 44 emotions while imagining the scenario.

We found that participants in the high risk scenario reported higher levels of negative individualistic emotions (e.g., angry, nervous), negative prosocial emotions (e.g., scornful, resentful) and lower levels of positive emotions (e.g., happy, grateful) than participants assigned to the medium and low risk scenarios. These differences were significant between high and low risk participants for both negative prosocial and negative individualistic emotions. These findings suggest that, unsurprisingly, use of safety-critical systems may involve strong negative emotions. The notion that computers are cognitive entities, with which interaction should be non-emotional in order to be efficient and successful, may be particularly destructive in this context. A lack of acknowledgment by the system may not only alter a user's decision-making, but lead to stronger negative emotions that impact later interaction.

Developing emotion-aware communication strategies by detecting users' emotions during system operations can reduce the potentially harmful effects of negative emotions. Specifically, teaching users to recognize their emotions (emotional education) may enable them to act more mindfully, and help to lessen the potential negative effects of strong emotions on decision-making (emotional inoculation) [6]. We argue that "emotional inoculation" is particularly applicable in the safety-critical domain, such as our hypothetical drone system. Communicating with users about emotions they may experience while using a system can positively contribute to both their decision-making outcomes and their perceptions of the system. Future work should test the effectiveness of safety-critical system interfaces that incorporate emotional inoculation via different types of messages and in various decision-making contexts. Furthermore, "emotional inoculation" and "emotional education" can be incorporated into training materials for safety-critical system users (e.g., drone operators). Using virtual simulators in realistic scenarios, such training systems could inform operators about the emotions they might experience during certain points of system use (e.g., feeling nervous and anxious during a time-sensitive task) and the nature of the specific emotions in such situations (e.g., prosocial vs. individualistic, or positive vs. negative). This can help prepare operators to regulate their reactions under time pressure and stress while performing complex safety-critical tasks [38].

These kinds of emotional communication can help to improve a user's trust calibration. Prior work has found that happiness, as well as "liking" a system influence reliance [39]. These affective aspects may help to explain changes in trust over the course of a human-computer interaction [29,39,40]. Future work should explore how the negative individualistic and negative prosocial emotions associated with safety-critical system use factor into trust evaluations and reliance decisions, as well as how an understanding of these emotions can be leveraged to improve system design and trust calibration.

We also found that, at the same risk level, the intensity of emotion factors differed (see Figs. 2 and 3). Negative prosocial emotions had the lowest mean intensity in all risk levels and roles, whereas positive emotions and negative individualistic emotions generally had higher intensity. Though prosocial emotions were not felt as strongly by participants, we observed that their anticipated intensity differed between high and low risk level participants. It appears that users are not just thinking about themselves with their use of the drone system, but about the involvement of others. This result is in line with research demonstrating the relevance of both individualistic and prosocial emotions in the context of pop-up security messages [22]. In the drone context, prosocial emotions could have been associated with (1) people on the ground who may have been impacted by the drone, (2) other human collaborators, or (3) the computer system itself. The latter is supported by research demonstrating social responses to computers by human users [24]. Future work could shed light on the specific effect that the computer itself has on user emotions by investigating how factors of the system and its interface influence the intensity of prosocial emotions, relative to differences in the context of system use.

Lastly, we found that for all the three factors, the interaction between risk level and role was not significant. This indicates that participants' emotions were more likely to be influenced by the criticality of the situation rather than their assigned role. It is possible that participants in operator and administrator roles in the same scenario considered the level of risk the same, and thus the role to which they were assigned did not make a strong contribution to their overall feelings. Such a difference may be more pronounced in a lab setting where participants interact directly with a system. If the user's role on a task-oriented team is more linked to their actions, then emotions may be impacted by their level of responsibility for team success.

5.1 Limitations

While this study provides insights about the effects of risk and role on users' emotions, there are several limitations in this work.

First, we used hypothetical (i.e., artificial) scenarios in which participants rated how they would expect to feel as the operator or administrator of a drone system. Given the lack of actual interaction with a computer system, it may have been difficult for participants to anticipate the emotions they would experience. Moreover, this could contribute to misinterpretations of the degree of risk. For example, some participants in the low risk condition (i.e., identifying whale pods) may have considered the situation to be very risky, since failure could have caused "job loss." Nevertheless, even in this artificial scenario-based methodology, our results revealed considerably diverse ratings of emotions depending on the group to which participants were assigned.

Second, we recruited participants from the MTurk platform. Although MTurk allows for recruiting larger and more diverse populations in terms of age, education level and ethnicity compared to samples from specific subpopulations (e.g., students enrolled in a psychology class) [41,42], it is hard to verify the attentiveness of MTurk users. To filter out responses that demonstrated a lack of understanding of the scenario, we included an attention check question in the survey.

Lastly, since our study was survey-based, emotional states of participants were measured via self-reports. Though our data provides insight into the role of "anticipated emotions" in a risky human-computer interaction, it needs further validation given that individuals may have difficulty predicting their emotional states [43]. To develop a more thorough understanding of user's emotions, future studies should investigate the somatic components (e.g., facial expressions and the heartbeat) [44] of "immediate emotions" in studies involving actual human interaction with a computer system.

We believe that this work is a useful starting point for research on the role of emotions in decision-making with safety-critical systems, which has important implications for system interface design. We encourage future work to investigate the specific factors that influence user emotions (e.g., risk and the nature of consequences, organizational structure, system features) as well as the influence that different types of emotions have on decision-making, behavior, and system performance.

6 Conclusion

This study aimed to understand the role of emotions in decisions at various risk levels and responsibilities with respect to a safety-critical system. Participants were asked to rate the intensity with which they would feel 44 emotions while imagining using a drone system in one of six hypothetical scenarios where they were asked to imagine themselves as a drone operator or system administrator in a high, medium, or low risk scenario. We found that participants assigned to the high risk scenario anticipated more intense negative individualistic, negative prosocial and less intense positive emotions than participants assigned to medium and low risk scenarios. We strongly believe that insights gained in this work will enable researchers to develop more effective emotionally-aware communication strategies for safety-critical systems.

Acknowledgments. This material is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-15-1-0490.

Appendix

The Descriptions of the Scenarios

Through a two-way 2 (role: system operator/system administrator) × 3 (risk level: high risk/medium risk/low risk) factorial design experiment, participants were assigned to one of the six hypothetical scenarios. Depending on the assigned role and risk level, participants were shown one of the phrasings separated by vertical bars (|) below. For instance, participants assigned to the system operator role in the high risk scenario were shown *($Role_{opr}$) and ($Risk_{high}$)*, while participants assigned to the system administrator role in the low risk scenario were shown *($Role_{adm}$) and ($Risk_{low}$)*, and so on. The entire written description of scenarios are outlined below.

Now, imagine that [$Role_{opr}$: *there are system administrators who are*] | [$Role_{adm}$: *you are the system administrator who is*] responsible for:

- Making sure that the software of the system that is used to operate the drone remotely is up-to-date.
- Making sure that the hardware of the system is up-to-date.
- Troubleshooting of the system if the performance is not acceptable.
- Performing preventative maintenance of the system.

However, despite [$Role_{opr}$: *their*] | [$Role_{adm}$: *your*] best effort, the system is not perfectly reliable and the system occasionally experiences the followings due to software bugs/hardware failures:

- The system occasionally crashes due to some unknown reasons and takes 2 min to reboot, making the system unavailable, and the timing/frequency of the crash is unpredictable.
- The system occasionally becomes very slow (e.g., freezes for 10 s at a time) due to unknown software/hardware bugs.
- The system occasionally drops video frames due to communication errors.
- Different hardware components of the system rarely fails (e.g., once every 6 months).

Now, imagine that [$Role_{opr}$: *you are asked to use the system to make*] | [$Role_{adm}$: *the drone that you are responsible for managing is going to be used by someone else (e.g., operator) whose*] decisions involve identifying:

- $Risk_{high}$: enemy targets in a battlefield where there may be also innocent civilians.
- $Risk_{med}$: arresting or not arresting suspected illegal immigrants who may be innocent citizens in a border region.
- $Risk_{low}$: whale pods or non-interesting seals in the ocean for a company.

Table 3. Participants were asked to rate the expected intensity of these 44 emotions on a scale ranging from 1 (the least amount of intensity) to 7 (the greatest amount of intensity).

Emotions
1. I would feel TRUSTING (e.g. because the system has given an opportunity to respond)
2. I would feel HAPPY (e.g., because I am informed of actual system states)
3. I would feel CONFIDENT (e.g., because I am informed of actual system states)
4. I would feel SECURE (e.g., because I am informed of actual system states)
5. I would feel SAD (e.g., because the system is not performing as expected)
6. I would feel DEPRESSED (e.g., because the system is not performing as expected)
7. I would feel DOWN (e.g., because the system is not performing as expected)
8. I would feel AFRAID (e.g., because the system is not performing as expected)
9. I would feel NERVOUS (e.g., because the system is not performing as expected)
10. I would feel ANXIOUS (e.g., because the system is not performing as expected)
11. I would feel ANGRY (e.g., because the system is not performing as expected)
12. I would feel INSULTED (e.g., because the system is not performing as expected)
13. I would feel HOSTILE (e.g., because the system is not performing as expected)
14. I would feel SURPRISED (e.g., because one does not expect the interruption)
15. I would feel DAZED (e.g., because one does not expect the interruption)
16. I would feel CONFUSED (e.g., because one does not expect the interruption)
17. I would feel FREAKED OUT (e.g., because one does not expect the interruption)
18. I would feel DISGUSTED (e.g., because the system is not performing as expected)
19. I would feel DISMAYED (e.g., because the system is not performing as expected)
20. I would feel DISTRAUGHT (e.g., because the system is not performing as expected)
21. I would feel CARED-FOR (e.g., because I am informed of actual system states)
22. I would feel FRIENDLY (e.g., because I am informed of actual system states)
23. I would feel WELCOMED (e.g., because I am informed of actual system states)
24. I would feel POWERFUL (e.g., because I am warned and can respond)
25. I would feel ENERGETIC (e.g., because I am warned and can respond)
26. I would feel VIGOROUS (e.g., because I am warned and can respond)
27. I would feel ISOLATED (e.g., because my response may be inadequate)
28. I would feel LONELY (e.g., because my response may be inadequate)
29. I would feel ABANDONED (e.g., because my response may be inadequate)
30. I would feel PROUD (e.g., because I am warned and can respond)
31. I would feel TRIUMPHANT (e.g., because I am warned and can respond)
32. I would feel ARROGANT (e.g., because I am warned and can respond)
33. I would feel ASHAMED (e.g., because my response may be inadequate)
34. I would feel GUILTY (e.g., because my response may be inadequate)
35. I would feel EMBARRASSED (e.g., because my response may be inadequate)
36. I would feel SCORNFUL (e.g., because the system state is fine)
37. I would feel CONTEMPTUOUS (e.g., because the system state is fine)
38. I would feel DISDAINFUL (e.g., because the system state is fine)
39. I would feel HUMILIATED (e.g., because the system state is fine)
40. I would feel DISHONORED (e.g., because the system state is fine)
41. I would feel RESENTFUL (e.g., because the system state is fine)
42. I would feel GRATEFUL (e.g., because the system has given an opportunity to respond)
43. I would feel RESPECTFUL (e.g., because the system has given an opportunity to respond)
44. I would feel ADMIRING (e.g., because the system has given an opportunity to respond)

References

1. Loewenstein, G.F., Weber, E.U., Hsee, C.K., Welch, N.: Risk as feelings. Psychol. Bull. **127**(2), 267 (2001)
2. Schlösser, T., Dunning, D., Fetchenhauer, D.: What a feeling: the role of immediate and anticipated emotions in risky decisions. J. Behav. Decis. Mak. **26**(1), 13–30 (2013)
3. Kahneman, D., Tversky, A.: Prospect theory: an analysis of decision under risk. Econom.: J. Econom. Soc. **47**, 263–291 (1979)
4. Loewenstein, G., Lerner, J.S.: The role of affect in decision making. Handb. Affect. Sci. **619**(642), 3 (2003)
5. Buck, R., Davis, W.A.: Marketing risk: emotional appeals can promote the mindless acceptance of risk. In: Roeser, S. (ed.) Emotions and risky technologies, pp. 61–80. Springer, Dordrecht (2010). https://doi.org/10.1007/978-90-481-8647-1_4
6. Buck, R., Ferrer, R.: Emotion, warnings, and the ethics of risk communication. In: Roeser, S., Hillerbrand, R., Sandin, P., Peterson, M. (eds.) Handbook of Risk Theory, pp. 693–723. Springer, Dordrecht (2012). https://doi.org/10.1007/978-94-007-1433-5_27
7. Lerner, J.S., Li, Y., Valdesolo, P., Kassam, K.S.: Emotion and decision making. Ann. Rev. Psychol. **66**, 799–823 (2015)
8. Wilson, R.A., Keil, F.C.: The MIT Encyclopedia of the Cognitive Sciences. MIT press, Cambridge (2001)
9. Harless, D.W., Camerer, C.F.: The predictive utility of generalized expected utility theories. Econom.: J. Econom. Soc. **62**, 1251–1289 (1994)
10. Clore, G.L.: Cognitive phenomenology: feelings and the construction of judgment. Const. Soc. Judgm. **10**, 133–163 (1992)
11. Clore, G.L., Schwarz, N., Conway, M.: Affective causes and consequences of social information processing. Handb. Soc. Cogn. **1**, 323–417 (1994)
12. Loewenstein, G.: Out of control: visceral influences on behavior. Organ. Behav. Hum. Decis. Process. **65**(3), 272–292 (1996)
13. Basso, M.R., Schefft, B.K., Ris, M.D., Dember, W.N.: Mood and global-local visual processing. J. Int. Neuropsychol. Soc. **2**(3), 249–255 (1996)
14. Conway, M., Giannopoulos, C.: Dysphoria and decision making: limited information use for evaluations of multiattribute targets. J. Pers. Soc. Psychol. **64**(4), 613 (1993)
15. Forgas, J.P.: Mood and judgment: the affect infusion model (AIM). Psychol. Bull. **117**(1), 39 (1995)
16. Knight, J.C.: Safety critical systems: challenges and directions. In: Proceedings of the 24th International Conference on Software Engineering, ICSE 2002, pp. 547–550. ACM, New York (2002)
17. Brave, S., Nass, C.: The Human-Computer Interaction Handbook, pp. 81–96. L. Erlbaum Associates Inc., Hillsdale (2003)
18. Brave, S., Nass, C., Hutchinson, K.: Computers that care: investigating the effects of orientation of emotion exhibited by an embodied computer agent. Int. J. Hum.-Comput. Stud. **62**(2), 161–178 (2005)
19. Picard, R.W., Klein, J.: Computers that recognise and respond to user emotion: theoretical and practical implications. Interact. Comput. **14**(2), 141–169 (2002)
20. Klein, J., Moon, Y., Picard, R.W.: This computer responds to user frustration: theory, design, and results. Interact. Comput. **14**(2), 119–140 (2002)

21. Jones, C., Jonsson, I.-M.: Using paralinguistic cues in speech to recognise emotions in older car drivers. In: Peter, C., Beale, R. (eds.) Affect and Emotion in Human-Computer Interaction. LNCS, vol. 4868, pp. 229–240. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-85099-1_20
22. Buck, R., Khan, M., Fagan, M., Coman, E.: The user affective experience scale: a measure of emotions anticipated in response to pop-up computer warnings. Int. J. Hum.-Comput. Interact. **34**, 1–10 (2017)
23. Bowles, S., Gintis, H.: Prosocial Emotions. The Economy As an Evolving Complex System III, pp. 339–366. Santa Fe Institute, Santa Fe (2005)
24. Reeves, B., Nass, C.I.: The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places. Cambridge University Press, Cambridge (1996)
25. Nass, C., Moon, Y.: Machines and mindlessness: social responses to computers. J. Soc. Issues **56**(1), 81–103 (2000)
26. Freedy, A., DeVisser, E., Weltman, G., Coeyman, N.: Measurement of trust in human-robot collaboration. In: 2007 International Symposium on Collaborative Technologies and Systems, CTS 2007, pp. 106–114. IEEE (2007)
27. Dzindolet, M.T., Peterson, S.A., Pomranky, R.A., Pierce, L.G., Beck, H.P.: The role of trust in automation reliance. Int. J. Hum.-Comput. Stud. **58**(6), 697–718 (2003)
28. Parasuraman, R., Riley, V.: Humans and automation: use, misuse, disuse, abuse. Hum. Factors **39**(2), 230–253 (1997)
29. Lee, J.D., See, K.A.: Trust in automation: designing for appropriate reliance. Hum. Factors **46**(1), 50–80 (2004)
30. Buck, R., Anderson, E., Chaudhuri, A., Ray, I.: Emotion and reason in persuasion: applying the ari model and the casc scale. J. Bus. Res. **57**(6), 647–656 (2004)
31. Kay, R.H., Loverock, S.: Assessing emotions related to learning new software: the computer emotion scale. Comput. Hum. Behav. **24**(4), 1605–1623 (2008)
32. Kaiser, H.F.: An index of factorial simplicity. Psychometrika **39**(1), 31–36 (1974)
33. Fabrigar, L.R., Wegener, D.T., MacCallum, R.C., Strahan, E.J.: Evaluating the use of exploratory factor analysis in psychological research. Psychol. Methods **4**(3), 272 (1999)
34. Raîche, G., Walls, T.A., Magis, D., Riopel, M., Blais, J.G.: Non-graphical solutions for Cattell's scree test. Methodol.: Eur. J. Res. Methods Behav. Soc. Sci. **9**(1), 23–29 (2013)
35. Stevens, J.P.: Applied Multivariate Statistics for the Social Sciences. Routledge, London (2012)
36. McKinley, R.K., Manku-Scott, T., Hastings, A.M., French, D.P., Baker, R.: Reliability and validity of a new measure of patient satisfaction with out of hours primary medical care in the united kingdom: development of a patient questionnaire. BMJ: Br. Med. J. **314**(7075), 193–198 (1997)
37. Robinson, J.P., Shaver, P.R., Wrightsman, L.S.: Criteria for scale selection and evaluation. Meas. Personal. Soc. Psychol. Attitudes **1**(3), 1–16 (1991)
38. Luini, L.P., Marucci, F.S.: Prediction-confirmation hypothesis and affective deflection model to account for split-second decisions and decision-making under pressure of proficient decision-makers. Cogn. Technol. Work **17**(3), 329–344 (2015)
39. Merritt, S.M.: Affective processes in human-automation interactions. Hum. Factors **53**(4), 356–370 (2011)
40. Merritt, S.M., Ilgen, D.R.: Not all trust is created equal: dispositional and history-based trust in human-automation interactions. Hum. Factors **50**(2), 194–210 (2008)

41. Chandler, J.J., Paolacci, G.: Lie for a Dime: when most prescreening responses are honest but most study participants are impostors. Soc. Psychol. Personal. Sci. **8**(5), 500–508 (2017)
42. Landers, R.N., Behrend, T.S.: An inconvenient truth: arbitrary distinctions between organizational, mechanical turk, and other convenience samples. Ind. Organ. Psychol. **8**(2), 142–164 (2015)
43. Picard, R.W., Picard, R.: Affective Computing, vol. 252. MIT press, Cambridge (1997)
44. Han, S., Lerner, J.S., Sander, D., Scherer, K.: Decision making. In: Sander, D., Scherer, K.R. (eds.) The Oxford Companion to Emotion and the Affective Sciences. Oxford University Press, Oxford (2009)

Anticipated Emotions in Initial Trust Evaluations of a Drone System based on Performance and Process Information

Theodore Jensen*[1], Mohammad Maifi Hasan Khan[1], Yusuf Albayram[1], Md Abdullah Al Fahim[1], Ross Buck[2], and Emil Coman[3]

[1]Department of Computer Science & Engineering, University of Connecticut
[2]Department of Communication, University of Connecticut
[3]Health Disparities Institute, University of Connecticut Health Center

Abstract

Trust in automation has been largely studied through a cognitive lens, though theories suggest that emotions play an important role. Understanding the affective aspects of human-automation trust can inform the design of systems that garner appropriate trust calibration. Toward this, we designed 4 videos describing a hypothetical drone system: one control, and three with additional performance or process information, or both. Participants reported the intensity of 19 emotions they would anticipate as system operator, perceptions of the system's trustworthiness, individual differences, and perceptions of the government law enforcement agency behind the system. We found that propensity to trust, risk-taking tendencies, and institutional trust influenced the intensity of anticipated emotions.[1]

Keywords— emotions; human-automation trust; perceived trustworthiness; initial trust

*Corresponding author. E-mail: theodore.jensen@uconn.edu.

[1]This paper extends Jensen et al. (in press) and reports on different data from the same study. Please contact the corresponding author for access to the paper.

1 Introduction

Due to the increasing complexity and variety of applications of automated systems, HCI researchers have investigated the phenomenon of a human's "trust" in automation (J. D. Lee & See, 2004; Dzindolet, Peterson, Pomranky, Pierce, & Beck, 2003). Inappropriate levels of trust in such systems can lead to *disuse* of reliable automation or *misuse* of unreliable automation, both of which have well-cited negative consequences (Parasuraman & Riley, 1997). To prevent these outcomes, calibration of trust has been suggested as a design goal—rather than trusting more, users should be able to "calibrate" their trust to an appropriate level given the reliability of the system (J. D. Lee & See, 2004).

Researchers have noted that human-automation trust has cognitive and affective components (J. D. Lee & See, 2004), suggesting that the affective components may be stronger indicators of trust (Madsen & Gregor, 2000). Despite this, studies of emotions and human-automation trust are relatively scarce. This can be problematic given increasing use of safety-critical automated systems (e.g., autonomous cars, medical diagnostic systems, unmanned aerial vehicles (Carlson, Desai, Drury, Kwak, & Yanco, 2014; Freedy, DeVisser, Weltman, & Coeyman, 2007)), where risks are likely to strongly influence users' emotions and decision-making (Loewenstein & Lerner, 2003). Understanding how emotions relate to trust in this context can not only help to improve trust calibration, but reveal the emotional effects of risky human-automation collaborations on users.

In the current study, we aimed to observe how various trust-related factors influenced emotions anticipated in the context of drone operation. We designed 4 videos to test the isolated and cumulative effects of performance and process information about a hypothetical drone system used to assist law enforcement in a surveillance task. We recruited 160 participants for an online survey, randomly assigned each to a video, and asked them to imagine being the system operator. Participants then responded to a survey on the anticipated intensity of 19 emotions, perceived trustworthiness of the drone system, and institutional trusting beliefs in situational normality and structural assurance. They also reported on

demographics, propensity to trust, and domain-specific risk-taking tendencies.

Emotions factored into hostility, positive, anxiety, and loneliness components and we constructed multiple linear regression models predicting each with system information, individual differences, and institutional trust. We found that financial risk-taking increased anticipated hostility emotions, while recreational risk-taking led to the anticipation of more intense positive and less intense anxiety emotions. Propensity to trust predicted less intense loneliness emotions. Institutional trust influenced emotions as well. Whereas greater perceptions of the institution's ability led to more intense hostility emotions, greater perceptions of the institution's benevolence led to less intense hostility. Perceptions of institutional integrity also appeared to increase positive and decrease anxiety emotions. Structural assurance led to the anticipation of less intense hostility and anxiety and more intense positive emotions. These results offer initial support for a relationship between human-automation trust and emotions, warranting future research on how operator emotions can be addressed to improve trust calibration.

2 Related Work

Imagine a busy intersection that you pass through on your way to work, where an automated system controls the traffic lights using sensors in the road. When the system functions reliably, it is almost invisible. The light turns from red to green and you move on. Yet imagine your frustration after being stuck indefinitely at a red light that won't change–you may curse at the incompetence of the technology and blame the system for wasted time. Moreover, you will likely adjust your perceptions of the system's trustworthiness, perhaps even avoiding this intersection in the future because of your frustration. This highlights the intersection between human-machine trust and emotions.

2.1 Trust in HCI

Prior work has not only explored trust between humans in online settings, social media and computer-mediated communication (Warner-Søderholm et al., 2018; McKnight, Choudhury, & Kacmar, 2002a; Walther & Bunz, 2005), but trust between a human and computers, robots, and automation in general (J. D. Lee & See, 2004; Hoff & Bashir, 2015). We adopt one of the most widely cited definitions of human-human trust, which has been employed previously in the HCI literature, from R. C. Mayer, Davis, and Schoorman (1995):

> The willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that party.

A key part of this definition is "vulnerability." A trustor necessarily accepts some risk in their decision to rely on the trustee. In the above example, *vulnerability* stems from the risk of wasting time at the light and the fact that the system's actions cannot be controlled. You may decide to avoid the intersection because you are no longer *willing to rely* on the system to give a green light in a reasonable amount of time.

R. C. Mayer et al. (1995) propose that perceptions of the trustee's ability, integrity, and benevolence (i.e., perceived trustworthiness characteristics or trusting beliefs) influence the trustor's willingness to rely. *Ability* or *competence* refers to the trustee's skills within the relevant domain. *Integrity* reflects a belief that the trustee acts based on a set of predictable principles. *Benevolence* relates the extent to which a trustee is motivated to do good for the trustor. These characteristics have been adapted in the human-automation domain as the *performance*, *process*, and *purpose*, respectively, of an automated trustee (J. Lee & Moray, 1992). *Performance* refers to the consistency and reliability of system behavior. *Process* details the qualities that govern system behavior, such as its algorithms. *Purpose* is the motive or goal of the system.

So, cognitively, your bad experience with the traffic light system informs your perceptions

of its trustworthiness and, thus, your expectations of its future behavior. Yet, at the same time, trust may be experienced emotionally. J. D. Lee and See (2004) highlight the importance of the affective aspect of trust in their influential review of the trust in automation literature: *"Ultimately trust is an affective response, but it is also influenced by analytic and analogical processes"*. Some researchers have begun to study the nature of system operators' and administrators' emotional experiences, as well as how information about a drone system is regarded by individuals in these roles (Albayram, Khan, Jensen, Buck, & Coman, 2018; Albayram, Jensen, Khan, Buck, & Coman, 2018). However, in general, the role of affect in human-automation trust is not well understood.

According to the Computers are Social Actors (CASA) paradigm, social science theories often extend to interactions where one person is replaced by a computer (Nass, Steuer, & Tauber, 1994). For example, one experiment found that people demonstrate politeness toward computers, in that they were less critical when giving a direct evaluation as opposed to feedback on a separate machine (Reeves & Nass, 1996) (see Nass et al. (1994) for more CASA examples). Thus, we look to the human-human trust literature for insights into the role of emotions in trust.

2.2 Emotion in Human-Human Trust

Though we refer distinctly to "emotions" and "trust," these words may represent constructs that are highly interrelated. Researchers have considered the relationship between trust and emotions in a number of ways.

Boone and Buck (2003) argue that emotional expression by humans functions as a marker for cooperative behavior and trustworthiness, citing several studies on the dyadic prisoner's dilemma. In this trust-based activity, each individual must decide either to cooperate with or defect from their partner. If both cooperate, they each serve a 2-year prison sentence. If both defect, they each serve 5 years. If one defects and one cooperates, the defector is free and the cooperator serves 10 years. As the act of cooperating makes you vulnerable

to the risk of your partner not cooperating, it is useful to know whether or not to trust your partner, wherein authors note the role of emotions: *"Being more emotionally expressive makes being trustworthy harder to fake and players would have greater confidence that what they see is the true state of the person in question"* (Boone & Buck, 2003). Although computers are not emotionally expressive in the way that humans are, the aforementioned CASA research suggests that human-computer trust too may involve "emotional" cues that signal the computer's trustworthiness to the human trustor, albeit differently from familiar human cues such as facial expressions and body language.

Applying the *affect-as-information* (Schwarz & Clore, 1983) and *affect infusion* models (Forgas, 1995), W. S. Lee and Selart (2012) argue that individuals' regulation of and attention to affect influence the extent to which emotions influence their trust. Specifically, *affect-as-information* suggests that individuals use their mood to inform their trust judgment. *Affect infusion*, also known as affect priming, suggests that trust judgments of atypical or peripheral objects (e.g., an unfamiliar target of initial trust) are more likely to be influenced by emotions than those using more direct cognitive processing. Authors point to the fact that perception of risk, an essential condition of trust, is influenced by emotions (W. S. Lee & Selart, 2012). This motivates the current investigation of the emotions involved in risky human-computer interaction.

Dunn and Schweitzer (2005) similarly investigated the influence of incidental emotions (i.e., those unrelated to the trustee) on trust evaluations in 5 experiments. They found that emotions with other-person control appraisals (e.g., anger) and those with weak control appraisals (happiness) had a greater influence on trust than emotions with personal (pride) or situational (sadness) control appraisals. They also found that awareness of the causes of emotions and familiarity with the trustee both lessened the impact of emotions on trust.

We focus on anticipated emotions and their relation to initial trustworthiness perceptions of an unfamiliar trustee. This initial trust has been also been termed "swift trust," distinguished from knowledge-based trust that builds based on observation of a trustee's

behavior (Meyerson, Weick, & Kramer, 1996). Given that this initial form of trust is not based on a wealth of information about the trustee, and is instead a category-based, heuristic process, it stands to reason that swift trust is influenced by emotions. Swift trust evaluations have been found to carry over to subsequent knowledge-based trust (Robert, Denis, & Hung, 2009), lending to the importance of understanding the emotional dynamics of initial trust evaluations.

Based on these works, the current study explores the various emotions individuals anticipate as a drone system operator in order to better understand the role of affect in initial trust in automated systems.

3 Methodology

In previous analysis of data from the current study (Jensen et al., in press), we found that information about the performance or process of a drone system led to increased perceptions of the its ability, while process information also increased perceptions of integrity. Also, financial risk-taking tendencies and perceptions of structural assurance increased perceptions of trustworthiness. The purpose of the current study is to elucidate the emotional aspects of this initial human-automation trust, toward which we posed the following research questions:

RQ1: What types of emotions do people anticipate as the operator of a safety-critical drone system?

RQ2: How are anticipated emotions influenced by trust-related factors?

RQ3: What is the relationship between anticipated emotions and initial perceived trustworthiness?

To investigate these questions, we created 4 different videos describing a hypothetical drone system. In a between-group study, we randomly assigned participants to watch one of the videos and subsequently asked them to respond to a survey imagining that they were the

operator of the drone system. By focusing on initial trust, we could observe how anticipated emotions relate to trust evaluations of the system based on the video alone, apart from factors such as experience with the system or its interface.

3.1 Design of Videos

We chose video as the mode of communication because it utilizes both visual and auditory information processing channels, leading to higher engagement (Tempelman-Kluit, 2006; Clark & Paivio, 1991; R. E. Mayer & Sims, 1994; Herron, York, Corrie, & Cole, 2006; Podszebka, Conklin, Apple, & Windus, 1998).

Narrated informational content was reviewed by authors over several iterations to ensure clarity and relevance to Lee and Moray's definitions (J. Lee & Moray, 1992) of *performance* and *process* applied to our drone system. The *purpose* of the system was given in all videos so that participants had sufficient information to understand the system and their role. Thus, the control group watched a video containing only "baseline" information (i.e., describing what the system was used for), and the three experimental groups watched a video containing the same baseline content followed by either performance or process information, or both. Table 1 shows the full narration transcript.

The video's visual content was taken from a publicly available video of drone operation. The original audio was replaced by narration recorded by one of the researchers. Videos were trimmed to the length of the narration to avoid video playing without narration as well as repeated visual content. Longer videos therefore contain visual content that shorter videos do not. Because the video displays neutral images of operators at a control panel, we expect that the narration describing a safety-critical drone task was more salient. However, we acknowledge the potential effect of differences in visual content and refer the reader to Table 2 to view the videos on YouTube. While participants watched clear versions of the videos, we blurred out certain parts in the shared links for privacy reasons.

3.2 Survey Structure and Measures

Before viewing the video, participants answered demographic questions on age, gender, computer proficiency, race, education, and military experience. After the video, they were shown the following text:

Now imagine that you are working for a law enforcement agency as the operator of the presented drone system. Your task is to identify, track, and neutralize the vehicles of human traffickers who could harm civilians if not detained. Please note that failure to identify violent criminals such as human traffickers can put innocent civilians' and law enforcement officers' lives at danger. Please answer the following questions assuming the presented operating conditions.

Participants were asked to reiterate the scenario in their own words to ensure that they understood their task and role as operator.

We first evaluated trusting beliefs in the system's *ability*, *integrity*, and *benevolence* (i.e., perceived trustworthiness characteristics) (R. C. Mayer et al., 1995). These items were adapted from McKnight, Choudhury, and Kacmar (2002b) and R. C. Mayer and Davis (1999) to refer to the drone system described in the video. The ability (4-item, $\alpha = 0.81$), integrity (5-item, $\alpha = 0.88$), and benevolence (5-item, $\alpha = 0.85$) sub-scales demonstrated good reliability.

Subsequently, participants answered questions about situational normality and structural assurance. These institutional trust items were adapted from Li, Hess, and Valacich (2008) and McKnight et al. (2002a) to refer to the government law enforcement agency in our hypothetical scenario. The 4-item structural assurance scale demonstrated excellent reliability ($\alpha = 0.95$), and each of the 3-item situational normality sub-scales had at least good reliability (ability $\alpha = 0.88$; integrity $\alpha = 0.94$; benevolence $\alpha = 0.87$). The trusting beliefs and initial trust items were all rated on 7-point Likert scales. All adapted scales are included in the

Appendix.

Next, participants reported on 19 emotions by answering the question, "How do you anticipate you would feel operating the drone system?" where each emotion item was phrased, "I would feel" followed by the emotion (e.g., "I would feel ashamed," "I would feel resentful") on a 7-point scale from "Not at all" to "Very much." The specific emotions were a subset of those studied in prior work on emotion in the HCI domain (Buck, Khan, Fagan, & Coman, 2018; Albayram, Jensen, et al., 2018; Albayram, Khan, et al., 2018).

Participants lastly reported on their propensity to trust other people using a 12-item, 5-point scale adapted directly from Frazier, Johnson, and Fainshmidt (2013) and risk-taking tendencies in 5 domains (financial, ethical, health/safety, recreational, social) using the 30-item, 7-point domain-specific risk-taking (DOSPERT) scale (Sitkin & Pablo, 1992; Weber, Blais, & Betz, 2002; Blais & Weber, 2006; Highhouse, Nye, Zhang, & Rada, 2017). The trust propensity scale had excellent reliability ($\alpha = 0.95$) and each of the 6-item risk-taking domain sub-scales had at least acceptable reliability (ethical $\alpha = 0.80$; financial $\alpha = 0.81$; health/safety $\alpha = 0.71$; recreational $\alpha = 0.81$; social $\alpha = 0.74$).

We included two manipulation check items at the end of the survey to validate that the system's performance and process were communicated in the videos. There were also two attention check questions throughout the survey to ensure thoughtful responses.

3.3 Recruitment

We posted the study as a Human Intelligence Task (HIT) on Amazon's Mechanical Turk (MTurk) service available to users 18 years or older, living in the United States, with at least 1000 completed HITs and a 95% HIT approval rating. When participants accepted the HIT, they were shown an information sheet and link connecting them to the study hosted on our university's Qualtrics server.

There were 3 pre-screening questions to prevent participants from guessing the eligibility criterion. Participants had to answer "No" to the question *"Have you ever operated drones*

in the past?". This screened out individuals having operated recreational drones in addition to systems like that described in our study, ensuring that the trustee was unfamiliar and that we observed *initial* trust. We did not disclose this eligibility criterion to any participant.

Ineligible participants were informed that they could not participate or be compensated. Eligible participants watched their assigned video and took the survey. At the end, these participants were given a code generated on Qualtrics to submit to MTurk for $3 of compensation. On average, the survey took participants 14.3 minutes (Median = 12.4 minutes, SD = 8.5 minutes).

The study was approved by our university's Institutional Review Board.

3.4 Sample Demographics

Of the 200 participants eligible after pre-screening, we removed the data of those who incorrectly answered at least one multiple choice attention check question, entered an ineligible age, or misunderstood the scenario based on their post-video reiteration. For the latter, participants who gave an unrelated response, referred to the "operator" in the third-person (i.e., suggesting that they were not imagining being the operator) or mentioned something not expressed in the video (e.g., "the military," "drug sales," "child sex traffickers") were removed from the data. Lastly, to ensure the video was fresh in participants' minds, we removed data of those who waited greater than 10 minutes after their video ended to advance to the next part of the survey. Ultimately, 163 were retained for analysis and balanced among video groups (see Table 2).

The sample consisted of 89 (54.6%) male and 74 (45.4%) female participants with ages ranging from 20 to 64 (Mean = 35.3, SD = 10.0). Regarding computer proficiency, 58 (35.6%) participants reported being "Competent," 82 (50.3%) "Proficient," and 23 (14.1%) "Expert." There were 123 white (75.5%), 18 African American (11.0%), 6 Hispanic (3.7%), 11 Asian (6.7%), 2 Native American and 3 other participants. Furthermore, 82.2% of participants reported having some post-secondary education at a college or university and 7

(4.3%) reported having served in the military.

Testing for demographic differences between video groups, we found no significant differences in terms of gender ($\chi^2(3) = 2.19$, $p = .53$). Moreover, Fisher's Exact Test revealed neither significant differences in terms of race ($p = .51$) nor military service ($p = .32$). Kruskal-Wallis tests demonstrated that groups were not significantly different in terms of age ($H(3) = 1.73$, $p = .63$), education level ($H(3) = 0.16$, $p = .98$), or computer proficiency ($H(3) = 2.60$, $p = .46$). Based on these results, we concluded that the four video groups were similar in terms of their demographics.

4 Evaluation

In order to better understand the emotions involved in participants' initial trust evaluations, we investigated how various trust-related factors influenced the 19 emotion items, and how these related to the perceived trustworthiness of the drone system.

4.1 Manipulation Check of Information Types Communicated in the Videos

First, to verify whether performance and process information were communicated in the narration, we included two manipulation check statements at the end of the survey:

1. I was made aware of the drone system's performance (i.e., how effective the system is about accomplishing its goal).

2. I was made aware of the drone system's process (i.e., how the system works to accomplish its goal).

Participants rated these two items on a 7-point Likert scale from "Strongly Disagree" to "Strongly Agree." We use Mann-Whitney U-tests to compare between participants who

received or did not receive a given type of information. We also report the effect size of U-tests using $r = Z/\sqrt{N}$ metric (Field, 2013).

Participants who received *performance* information in their video (i.e., *Performance* and *Perf-Proc* groups) rated their awareness of the drone system's performance higher than other participants (i.e., *Control* and *Process* groups), though this difference was only marginally significant ($U = 2841.00$; $p = .10$; $r = -.13$). It may be that because the *performance* information mentioned potential system errors, these participants actually felt somewhat unaware of the system's performance.

Participants who received *process* information in their video (i.e., *Process* and *Perf-Proc* groups) rated their awareness of the system's process significantly higher than other participants (i.e., *Control* and *Performance* groups) ($U = 2667.50$; $p = .02$; $r = -.18$).

4.2 Factor Analysis of Emotion Items

Next, we conducted Principal Components Analysis (PCA) to observe whether the 19 emotion items aligned into the same factors as in Buck et al. (2018). We first used diagnostic tests to ensure that PCA was appropriate for the emotion survey items. The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy was 0.906, above the 0.9 referred to as a "marvelous" indicator of factorial simplicity (Kaiser, 1974). Bartlett's test of sphericity was significant ($\chi^2(171) = 2367.9$, $p < 0.001$), while all items had Spearman's correlations of greater than 0.5 with at least one other item. Thus, we determined that the 19 items were suited for reduction into components.

Using orthogonal Varimax rotation, our PCA revealed 4 factors, roughly aligning with the hostility, positive, anxiety, and loneliness factors in Buck et al. (2018). Two of the four items from Buck et al.'s loneliness factor that we included in our survey ("I would feel ashamed," "I would feel humiliated") loaded with the hostility factor, while the remaining items factored as expected. These 4 factors accounted for 75.4% of the variance in anticipated emotion ratings (hostility 44.3%, positive 16.6%, anxiety 8.4%, and loneliness 6.1%). Factor loadings

for each item, as well as Cronbach's α and the average inter-item correlations (IIC) for each factor are shown in Table 3. Each factor score used in subsequent analyses was computed by averaging a participants' ratings for its subsisting emotion items.

4.3 Multiple Linear Regressions: What Trust-Related Factors influence Anticipated Emotions?

As prior work on trust in automation has suggested an affective component of trust, we investigated how the trust-related factors we measured influenced the emotion factors.

We built a separate regression model predicting each emotion factor score based on individual differences (trusting propensity, risk-taking domains), system information (performance and process information), and institutional trust (structural assurance, as well as situational normality perceptions of the ability, integrity and benevolence of government law enforcement agencies). These align with the dispositional, learned, and situational dimensions of trust in automation proposed by Hoff and Bashir (2015). Performance and process dummy variables respectively indicate whether performance or process information was present in a participants' video. Variance inflation factors were highest for the institutional trust predictors, though all were less than 7.5, indicating that collinearity was not a major concern. Multiple linear regression models for each of the 4 emotion factors are displayed in Table 4.

Although the experimental manipulations (i.e., system information contributing to learned trust) did not appear to influence anticipated emotions, there were significant effects due to dispositional and situational factors.

4.3.1 Dispositional Factors: Trusting Propensity and Risk-Taking Tendencies

Participants with a greater propensity to trust were likely to anticipate less intense loneliness (β = -0.28, p< 0.05), suggesting that individuals with low levels of trust may experience feelings of isolation.

Financial risk-takers were likely to anticipate slightly greater hostility emotions (β = 0.14, $p< 0.05$). Recreational risk-takers were likely to anticipate more intense positive (β = 0.19, $p< 0.05$) and less intense anxiety emotions (β = -0.24, $p< 0.01$), reflecting a greater sense of comfort with the risks in the safety-critical drone task.

4.3.2 Situational Factors: Structural Assurance and Situational Normality

Greater perceptions of law enforcement agencies' ability were associated with more intense hostility emotions ($\beta = 0.39$, $p< 0.05$), while greater perceptions of benevolence were associated with less intense hostility (β = -0.32, $p< 0.05$). While the latter result is predictable, as those who view the institution behind the system as benevolent are likely to anticipate being less "scornful" or "ashamed," the former was more surprising. This result may reflect a distaste for the use of drones in the surveillance context, such that the perceived competence of law enforcement agencies prompted resentment towards them.

Increased perceptions of the integrity of law enforcement agencies were associated with the anticipation of more intense positive ($\beta = 0.35$, $p< 0.05$) and less intense anxiety emotions (β = -0.49, $p< 0.01$). This finding shows that perceptions of institutional trustworthiness and reliability may positively influence a system user's emotional experience.

4.4 How do Emotions Relate to Trust?

Lastly, we performed non-parametric correlation analysis (Spearman's ρ) to observe the associations between anticipated emotions and the perceived trustworthiness of the drone system. Correlations between each emotion factor and trusting belief are shown in Table 5. The strongest correlations are those between the trusting beliefs and positive emotions—perceived ability had the strongest association ($\rho = 0.48$, $p< 0.01$), followed by integrity (ρ = 0.36, $p< 0.01$) and benevolence ($\rho = 0.29$, $p< 0.01$). Participants' positive perceptions of the drone system's trustworthiness were associated with anticipating being more "happy" and "secure," while the perceived competence of the system in this task was a slightly stronger

emotional indicator than its perceived consistency or whether it was concerned about the well-being of the operator. Perceived ability was negatively correlated with hostility (ρ = -0.26, p< 0.01), anxiety (ρ = -0.30, p< 0.01), and loneliness (ρ = -0.28, p< 0.01) emotions, as was perceived integrity with hostility (ρ = -0.19, p< 0.05), anxiety (ρ = -0.17, p< 0.05), and loneliness (ρ = -0.16, p< 0.05), but to a lesser extent.

5 Discussion

Our findings suggest that a system operator's emotional experience is indeed related to their trust. Although we are unable to determine whether a directed causal relationship exists (e.g., higher perceptions of trustworthiness lead to more positive emotions), these findings confirm the viability of an affective component to trust, whereas most prior work on trust in automation has focused on cognitive trust evaluations.

Regarding **RQ1**, we found that hostility, positive, anxiety, and loneliness emotion components explained 75.4% of the variance in anticipated emotion ratings in the current study. "Ashamed" and "humiliated," two items that loaded with loneliness in the study of pop-up software update warnings (Buck et al., 2018) were associated with emotions such as "scornful" and "resentful" in this case. This suggests that hostility regarding the use of drones in the presented surveillance context was closely associated with personal feelings of shame, distinctly from a sense of isolation and loneliness experienced by participants.

Mean ratings of hostility (M = 1.62, SD = 1.04) and loneliness (M = 1.96, SD = 1.27) indicate that they were not anticipated to be as intense as positive (M = 4.32, SD = 1.55) or anxiety (M = 2.65, SD = 1.34) emotions. The high degree of anticipated positive emotions suggests that the description of the scenario did not lead participants to anticipate feeling overtly negative, although a degree of anxiety was present given the safety-critical nature of the task. Nonetheless, even the mean rating for positive emotions is only slightly above the midpoint of the scale, which may reflect a difficulty in both reporting on one's own emotions

and in anticipating feelings in the hypothetical scenario.

For **RQ2**, as in our prior regression analysis of factors influencing the perceived ability, integrity, and benevolence of the drone system (Jensen et al., in press), we found that dispositional and situational factors influenced the degree of anticipated emotions.

Trusting propensity (e.g., "Trusting another person is not difficult for me," "I believe that people usually keep their promises") was only a significant predictor of loneliness, with a greater propensity to trust leading to the anticipation of less intense loneliness. This suggests that feeling isolated may be symptomatic of low trust. We encourage future work to observe how variations in system trustworthiness affect operator loneliness, and whether communication by the system can temper this feeling for those with a lesser propensity to trust.

Also, while we had previously found that financial risk-taking (e.g., "Betting a day's income at a high-stakes poker game") was associated with more positive trustworthiness perceptions, it positively predicted hostility in the current analysis. Perhaps, the impersonal nature of financial risks means financial risk-takers are more likely to report being "hostile" or "ashamed" with respect to the other people in this context. This also contrasts with the finding that recreational risk-takers (e.g., "Going down a ski run that is beyond your ability") were likely to anticipate less intense anxiety and more intense positive emotions, a predictable result of their comfort in the safety-critical scenario. These findings support the proposition by Lee and Selart that individual differences in regulation of affect, a construct certainly related to risk-taking tendencies, determine the extent to which emotions influence trust judgments (W. S. Lee & Selart, 2012).

Our previous finding that the perception of structural assurance positively influenced perceived trustworthiness was corroborated by the current analysis, in that structural assurance led to the anticipation of less intense hostility and anxiety and more intense positive emotions. On top of this, while we had found that situational normality (i.e., perceptions of law enforcement agencies' trustworthiness in this context) did not appear to influence

the perceived trustworthiness of the drone system, it did influence the degree of anticipated emotion. Greater perceptions of the integrity of law enforcement agencies (e.g., "I am comfortable relying on government law enforcement agencies to meet their obligations") were likely to reduce the intensity of anticipated anxiety and increase that of positive emotions. Likewise, situational normality benevolence perceptions (e.g., "If a system operator required help, most government law enforcement agencies would do their best to help") were likely to decrease the anticipated intensity of hostility emotions. Somewhat surprisingly, situational normality ability perceptions (e.g., "I feel that most government law enforcement agencies are good at what they do") appeared to increase the anticipated intensity of hostility emotions. This speaks to the fact that, in this context, emotions may not be associated simply with perceptions that the law enforcement agency is trustworthy, but with concerns about the morality of drone use for surveillance. In this sense, a more capable law enforcement agency is a more threatening one. This ethical influence is supported by the positive effect of situational normality benevolence.

None of the factors of learned trust (i.e., system information in the video) appeared to influence the intensity of anticipated emotions. This contrasts with our prior findings that performance and process information influenced the perceived ability and integrity of the drone system, and may speak to the distinction between cognitive and affective aspects of trust. Whereas trusting beliefs are more knowledge-based, affect appears to be associated more with the category-based processing that defines swift trust (Robert et al., 2009). This is also supported by the role of situational normality perceptions in anticipated emotion—the degree to which participants perceived institutions in the same category to be trustworthy influenced emotions, but did not necessarily inform perceptions of the trustworthiness of the system. Future work adapting Dunn and Schweitzer's study (Dunn & Schweitzer, 2005) to observe the influence of incidental emotions as well as institutional perceptions on initial trust and subsequent trust development in an automated system would shed light on the nature of swift and knowledge-based trust in HCI.

Lastly, toward answering **R3**, we found that the emotion factors were moderately correlated with the perceived trustworthiness characteristics of the drone system. Positive emotions demonstrated the strongest associations with the trusting beliefs. Moreover, correlation coefficients were strongest for the ability belief, followed by integrity, and finally by benevolence, where some even fell out of significance. This is in line with prior human-human trust research finding that ability perceptions have a greater influence on trust than integrity, which has a greater influence than benevolence (Robert et al., 2009). This also suggests that emotional intensity could be an indicator of a system operator's trust level, or that a positive emotional experience increases trust.

We stress that, because appropriate trust calibration is a desirable goal, designers should avoid attempts at simply increasing trust, even if the result is a less positive emotional experience for system operators. Instead, knowledge of how operators' emotions relate to their trust can help in recognizing inappropriate fluctuations in perceived trustworthiness of a system. This can in turn be applied to avoid the negative, potentially dangerous consequences of both undertrust and overtrust.

6 Conclusion

This study sheds light on the emotional aspects of a human-machine trust scenario. In order to observe how anticipated emotions relate to initial trust perceptions, participants were introduced to a drone system with one of four narrated videos. Although information in the videos did not appear to have an effect on emotions, we found that variation in the anticipation of hostility, positive, anxiety, and loneliness emotions was explained by dispositional and situational trust factors that were previously found to influence perceptions of the drone system's trustworthiness. Perceptions of the ability, integrity, and benevolence of the institution behind the system also influenced the intensity of emotions anticipated as operator. Lastly, we found moderate correlations between emotions and the perceived

trustworthiness of the system, with positive emotions having the strongest associations with trusting beliefs. We encourage future work to build upon this exploratory study in order to characterize the relationship between human-machine trust and emotions in more depth. This can improve our understanding of how human-machine trust environments emotionally impact system operators, as well as how emotions impact their trust, which can aid the design of systems that promote appropriate trust calibration.

7 Acknowledgments

This material is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-15-1-0490.

References

Albayram, Y., Jensen, T., Khan, M. M. H., Buck, R., & Coman, E. (2018). Investigating the effect of system reliability, risk, and role on users' emotions and attitudes toward a safety-critical drone system. *International Journal of Human–Computer Interaction*, 1–12.

Albayram, Y., Khan, M. M. H., Jensen, T., Buck, R., & Coman, E. (2018). The effects of risk and role on users anticipated emotions in safety-critical systems. In *International conference on engineering psychology and cognitive ergonomics* (pp. 369–388).

Blais, A.-R., & Weber, E. U. (2006). A domain-specific risk-taking (dospert) scale for adult populations.

Boone, R. T., & Buck, R. (2003). Emotional expressivity and trustworthiness: The role of nonverbal behavior in the evolution of cooperation. *Journal of Nonverbal Behavior*, *27*(3), 163–182.

Buck, R., Khan, M., Fagan, M., & Coman, E. (2018). The user affective experience scale: A measure of emotions anticipated in response to pop-up computer warnings. *International Journal of Human–Computer Interaction*, *34*(1), 25–34.

Carlson, M. S., Desai, M., Drury, J. L., Kwak, H., & Yanco, H. A. (2014). Identifying factors that influence trust. In *2014 aaai spring symposium series.*

Clark, J. M., & Paivio, A. (1991). Dual coding theory and education. *Educational psychology review*, *3*(3), 149–210.

Dunn, J. R., & Schweitzer, M. E. (2005). Feeling and believing: the influence of emotion on trust. *Journal of personality and social psychology*, *88*(5), 736.

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.

Field, A. (2013). *Discovering statistics using ibm spss statistics*. sage.

Forgas, J. P. (1995). Mood and judgment: the affect infusion model (aim). *Psychological bulletin*, *117*(1), 39.

Frazier, M. L., Johnson, P. D., & Fainshmidt, S. (2013). Development and validation of a propensity to trust scale. *Journal of Trust Research*, *3*(2), 76–97.

Freedy, A., DeVisser, E., Weltman, G., & Coeyman, N. (2007). Measurement of trust in human-robot collaboration. In *Collaborative technologies and systems, 2007. cts 2007. international symposium on* (pp. 106–114).

Herron, C., York, H., Corrie, C., & Cole, S. P. (2006). A comparison study of the effects of a story-based video instructional package versus a text-based instructional package in the intermediate-level foreign language classroom. *Calico Journal*, 281–307.

Highhouse, S., Nye, C. D., Zhang, D. C., & Rada, T. B. (2017). Structure of the dospert: Is there evidence for a general risk factor? *Journal of Behavioral Decision Making*, *30*(2), 400–406.

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on

factors that influence trust. *Human Factors*, *57*(3), 407–434.

Jensen, T., Albayram, Y., Khan, M. M. H., Buck, R., Coman, E., & Fahim, M. A. A. (in press). Initial trustworthiness perceptions of a drone system based on performance and process information. In *Int. Conference on Human Agent–Interaction (HAI '18)*. doi: 10.1145/3284432.3284435

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika*, *39*(1), 31–36.

Lee, J., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, *35*(10), 1243–1270.

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

Lee, W. S., & Selart, M. (2012). The impact of emotions on trust decisions.

Li, X., Hess, T. J., & Valacich, J. S. (2008). Why do we trust new technology? a study of initial trust formation with organizational information systems. *The Journal of Strategic Information Systems*, *17*(1), 39–71.

Loewenstein, G., & Lerner, J. S. (2003). The role of affect in decision making. *Handbook of affective science*, *619*(642), 3.

Madsen, M., & Gregor, S. (2000). Measuring human-computer trust. In *11th australasian conference on information systems* (Vol. 53, pp. 6–8).

Mayer, R. C., & Davis, J. H. (1999). The effect of the performance appraisal system on trust for management: A field quasi-experiment. *Journal of applied psychology*, *84*(1), 123.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of management review*, *20*(3), 709–734.

Mayer, R. E., & Sims, V. K. (1994). For whom is a picture worth a thousand words? extensions of a dual-coding theory of multimedia learning. *Journal of educational psychology*, *86*(3), 389.

McKnight, D. H., Choudhury, V., & Kacmar, C. (2002a). Developing and validating trust

measures for e-commerce: An integrative typology. *Information systems research*, *13*(3), 334–359.

McKnight, D. H., Choudhury, V., & Kacmar, C. (2002b). The impact of initial consumer trust on intentions to transact with a web site: a trust building model. *The journal of strategic information systems*, *11*(3-4), 297–323.

Meyerson, D., Weick, K. E., & Kramer, R. M. (1996). Swift trust and temporary groups. *Trust in organizations: Frontiers of theory and research*, *166*, 195.

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 72–78).

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

Podszebka, D., Conklin, C., Apple, M., & Windus, A. (1998). Comparison of video and text narrative presentations on comprehension and vocabulary acquisition.

Reeves, B., & Nass, C. I. (1996). *The media equation: How people treat computers, television, and new media like real people and places.* Cambridge university press.

Robert, L. P., Denis, A. R., & Hung, Y.-T. C. (2009). Individual swift trust and knowledge-based trust in face-to-face and virtual team members. *Journal of Management Information Systems*, *26*(2), 241–279.

Schwarz, N., & Clore, G. L. (1983). Mood, misattribution, and judgments of well-being: informative and directive functions of affective states. *Journal of personality and social psychology*, *45*(3), 513.

Sitkin, S. B., & Pablo, A. L. (1992). Reconceptualizing the determinants of risk behavior. *Academy of management review*, *17*(1), 9–38.

Tempelman-Kluit, N. (2006). Multimedia learning theories and online instruction. *College & Research Libraries*, *67*(4), 364–369.

Walther, J. B., & Bunz, U. (2005). The rules of virtual groups: Trust, liking, and performance in computer-mediated communication. *Journal of communication*, *55*(4), 828–846.

Warner-Søderholm, G., Bertsch, A., Sawe, E., Lee, D., Wolfe, T., Meyer, J., ... Fatilua, U. N. (2018). Who trusts social media? *Computers in Human Behavior*, *81*, 303–315.

Weber, E. U., Blais, A.-R., & Betz, N. E. (2002). A domain-specific risk-attitude scale: Measuring risk perceptions and risk behaviors. *Journal of behavioral decision making*, *15*(4), 263–290.

8 Appendix

8.1 Adapted Survey Items

8.1.1 Perceived Trustworthiness Characteristics (Trusting Beliefs)

Adapted from McKnight et al. (2002b), originally from R. C. Mayer and Davis (1999). Rated on a 7-point Likert scaled from "Strongly Disagree" to "Strongly Agree."

- **Ability**
  - The drone system would be competent and effective at assisting in tracking enemy targets.
  - The drone system would perform its role of neutralizing enemy targets very well.
  - Overall, the drone system would be a capable and proficient means for stopping the targets.
  - In general, the drone system would be very knowledgeable about stopping criminals.
- **Integrity**
  - The drone system would be truthful in its communication with me.
  - I would characterize the drone system as honest.

- The drone system would keep its commitments.
- The drone system would be sincere and genuine.
- The drone system would perform as expected.

- **Benevolence**
  - I believe that the drone system would operate in my best interest.
  - If I required help, the drone system would do its best to help me.
  - The drone system would be concerned about my well-being, not just its own.
  - The drone system would be concerned about the well-being of officers on the ground.
  - The drone system would be concerned about the well-being of civilians.

8.1.2 Institutional Trust

Adapted from Li et al. (2008) and McKnight et al. (2002a). Rated on a 7-point Likert scale from "Strongly Disagree" to "Strongly Agree."

- **Structural Assurance**
  - Government law enforcement agencies have enough safeguards to make me feel comfortable using drone systems.
  - I feel assured that legal and technological structures within the government law enforcement agencies would adequately protect me from problems while using the drone system.
  - I feel confident that technological advances within the government law enforcement agencies make it safe for me to use the drone system.
  - In general, government sponsored software systems are robust and safe to use.

- **Situational Normality - *Ability***

  - In general, most government law enforcement agencies are competent at developing software.
  - Most government law enforcement agencies do a capable job at meeting the needs of their system operators.
  - I feel that most government law enforcement agencies are good at what they do.

- **Situational Normality - *Integrity***
  - I am comfortable relying on government law enforcement agencies to meet their obligations.
  - I feel fine using government software systems since government law enforcement agencies generally fulfill their agreements.
  - I always feel confident that I can rely on government law enforcement agencies when I interact with their software systems.

- **Situational Normality - *Benevolence***
  - I feel that most government law enforcement agencies would act in a system operator's best interest.
  - If a system operator required help, most government law enforcement agencies would do their best to help.
  - Most government law enforcement agencies are interested in the well-being of their systems' users, not just their own well-being.

8.2 Tables

Baseline	Hello! The video you are watching presents a hypothetical scenario where operators are using a drone system to assist government law enforcement in stopping human traffickers. The system consists of an Unmanned Aerial Vehicle, or UAV, and a display that the operator observes while controlling the system. The operator is responsible for navigation of the drone and reporting locations of suspected human traffickers. Timely and accurate identification of violent criminals is extremely important as failures can put innocent civilians' and law enforcement officers' lives in danger. While the operators may shoot at targets if necessary, this is used only as the last option, as it could lead to hitting innocent civilians near the target or causing property damage.
Performance	While the system operates effectively most of the time, there can be occasional errors that impact video quality and drone maneuverability, caused by factors such as poor network connections and software glitches. As a result, operators may experience rare events such as screen blackouts or loss of connectivity lasting at most a few seconds.
Process	To make the system robust against such failures, the UAV has on-board algorithms that use sensors to improve flight stability and maneuverability. Information about system health is also automatically monitored and sent back to the operator over a network connection. This allows the operator to monitor and override system control if needed.

Table 1: Narration script for the videos. All videos contained "baseline" information describing the purpose of the system, while the *Performance* video additionally contained the performance information, *Process* the process information, and *Perf-Proc* both types of additional system information.

Label	n	Link to Video	Length
Control	41	`https://youtu.be/DuMwSsrEG5s`	50 s
Performance	39	`https://youtu.be/RJdwtSuGmAc`	71 s
Process	39	`https://youtu.be/2BTbNTAG19A`	70 s
Perf-Proc	44	`https://youtu.be/c5JrIdQNkY4`	92 s

Table 2: List of the 4 videos used in the study, which can be viewed on YouTube. The original video can be found at: `https://www.dvidshub.net/video/411919/mq1b-predator-gcs-broll`.

	Hostility	Positive	Anxiety	Loneliness
Disdainful	**0.884**			
Scornful	**0.857**			
Contemptuous	**0.851**			
Hostile	**0.774**			
Resentful	**0.865**			
Ashamed	**0.740**			
Humiliated	**0.702**			
Confident		**0.806**		
Secure		**0.798**		
Grateful		**0.842**		
Happy		**0.815**		
Respectful		**0.843**		
Nervous			**0.810**	
Anxious			**0.840**	
Confused			**0.700**	
Afraid	0.472		**0.665**	
Freaked out	0.409		**0.660**	
Lonely				**0.865**
Isolated				**0.855**
α	0.936	0.904	0.889	0.851
IIC	0.677	0.656	0.619	0.744

Table 3: Factor loadings for each of the 19 emotion items. Loadings less than 0.4 are not shown. The highest loading for each emotion is shown in bold. Scale reliability with Cronbach's α and the average inter-item correlations for each factor are shown in the last two rows.

	Hostility			Positive			Anxiety			Loneliness		
Predictors	β	SE	p	β	SE	p	β	SE	p	β	SE	p
Individual Differences												
Trusting propensity	0.11	0.09	0.24	0.06	0.11	0.61	-0.19	0.12	0.12	-0.28	0.13	**0.03**
Risk-taking:												
Ethical	0.11	0.08	0.21	-0.16	0.10	0.12	0.04	0.11	0.70	0.13	0.11	0.26
Financial	0.14	0.06	**0.03**	0.11	0.08	0.14	-0.04	0.08	0.67	0.14	0.09	0.12
Health/safety	0.11	0.08	0.18	0.02	0.10	0.83	0.16	0.11	0.12	0.02	0.11	0.88
Recreational	-0.07	0.07	0.28	0.19	0.08	**0.02**	-0.24	0.09	**0.01**	0.02	0.09	0.83
Social	-0.07	0.07	0.33	-0.06	0.08	0.46	-0.04	0.09	0.69	-0.15	0.09	0.11
System Information												
Perf	-0.25	0.20	0.21	0.10	0.24	0.66	< 0.01	0.26	≈ 1.00	-0.38	0.27	0.16
Proc	-0.01	0.20	0.95	0.13	0.24	0.58	0.01	0.26	0.97	-0.43	0.27	0.11
Perf x Proc	0.18	0.27	0.52	0.22	0.33	0.52	-0.35	0.36	0.33	0.39	0.38	0.31
Institutional Trust												
Structural Assurance	-0.29	0.13	**0.03**	0.39	0.15	**0.01**	-0.33	0.17	**0.05**	-0.17	0.18	0.35
Situational Normality:												
Ability	0.39	0.14	**0.01**	0.08	0.17	0.64	0.35	0.19	0.06	0.01	0.19	0.94
Integrity	-0.14	0.12	0.26	0.35	0.15	**0.02**	-0.49	0.16	$<$ **0.01**	-0.22	0.17	0.20
Benevolence	-0.32	0.14	**0.02**	-0.01	0.17	0.95	0.13	0.18	0.49	0.19	0.19	0.32
Constant	2.85	0.52	$<$ **0.01**	-0.39	0.63	0.54	5.39	0.69	$<$ **0.01**	4.12	0.72	$<$ **0.01**
Adjusted R^2	0.3109			0.5480			0.2873			0.1359		
$F(13, 149) =$	6.62 ($p < .001$)			16.11 ($p < .001$)			6.02 ($p < .001$)			2.96 ($p < .001$)		

Table 4: Results of the four separate multiple linear regressions, each predicting an emotion factor derived from PCA based on the various predictors. p-values which are significant at the 0.05 level are shown in bold.

	Hostility	**Positive**	**Anxiety**	**Loneliness**
Ability	**-0.258**	**0.479**	**-0.296**	**-0.284**
Integrity	**-0.189**	**0.361**	**-0.168**	**-0.155**
Benevolence	-0.005	**0.290**	-0.072	0.021

Table 5: Spearman's ρ correlations between the trusting beliefs (i.e, perceived trustworthiness of the drone system) and anticipated emotion factor scores. Coefficients significant at the 0.05 level are shown in bold.

Effect of Feedback and Warning Reliability on Trust, Emotions, and System Usage

Md Abdullah Al Fahim
Department of Computer Science & Engineering
University of Connecticut
Storrs, United States
md.fahim@uconn.edu

Mohammad Maifi Hasan Khan
Department of Computer Science & Engineering
University of Connecticut
Storrs, United States
maifi.khan@uconn.edu

Yusuf Albayram
Department of Computer Science & Engineering
University of Connecticut
Storrs, United States
yusuf.albayram@uconn.edu

Theodore Jensen
Department of Computer Science & Engineering
University of Connecticut
Storrs, United States
theodore.jensen@uconn.edu

Ross Buck
Communication and Psychological Sciences
University of Connecticut
Storrs, United States
ross.buck@uconn.edu

Emil Coman
Health Disparities Institute
University of Connecticut
Health Center Farmington,
United States
coman@uchc.edu

ABSTRACT
Safety-critical systems (e.g., UAV) are often equipped with warning mechanisms to alert users of imminent hazards. However, they can suffer from false alarms, and affect users' emotions and trust in the system negatively. While providing feedback could be an effective way to repair trust under such scenarios, the effects of warning reliability and feedback on users' emotions, trust, and behavior is not clear. This paper attempts to address this void by designing a 2 (warning reliability) x 2 (feedback) between-group study where participants interacted with a simulated UAV system to identify and neutralize enemy targets. Results indicated that feedback negatively affected users' positive emotions and trust in the system, and increased negative emotions. While emotions were found to mediate the relationship between feedback and trust, however, compliance behavior was not affected by trust. Implications of our findings for designing feedback and calibration of trust are discussed in the paper.

Author Keywords
Warning; Trust; Emotion; Compliance; Feedback; Safety-critical systems; Trust Mediation

CCS Concepts
•**Human-centered computing** → **Empirical studies in HCI; User centered design;** *Graphical user interfaces; Auditory feedback;* Please use the 2012 Classifiers and see this link to embed them in the Tex: https://dl.acm.org/ccs/ccs_flat.cfm

CHI'16, May 07–12, 2016, San Jose, CA, USA

ISBN 123-4567-24-567/08/06…$15.00

DOI: http://dx.doi.org/10.475/123_4

INTRODUCTION
Widespread adoption of wired and wireless sensor technologies has led to the emergence of various dynamic data-driven safety-critical decision support systems (e.g., battlefield monitoring, target tracking). Although such systems are designed with great care, they can still fail due to various reasons such as hardware/software failures, environmental factors, misconfiguration, and/or human errors, among other factors [2]. Unfortunately, sudden exposure to such system errors can negatively affect users' decision-making abilities and priority assessment of the subsequent tasks [35], limiting the capacity to process information mindfully [67]. To minimize such negative effects due to unexpected system malfunctions, safety-critical systems often incorporate warning modules that alert users regarding imminent hazards (e.g., system failures, environmental conditions). However, due to the nondeterministic nature of error probability, these warning systems are often not perfect, and trigger false alarms [44, 50, 77].

As it has been shown that users apply social norms while interacting with computers [51, 53, 54], it is likely that such false alarms may evoke strong negative emotions, which is shown to influence cognition, decision-making, and subsequent actions (i.e., commonly known as 'control precedence' [29]) [32], and users' trust in the system and the decision making process negatively [37].

As the effect of system errors is shown to be instantaneous [83], we hypothesize that providing real-time feedback can be an effective way to counter negative emotions caused by false alarms and system errors, and foster positive emotions. To better understand the effects of feedback and system reliability on users' emotions, trust, and response behavior, we designed

a 2 (warning reliability) x 2 (feedback) between-group study where participants interacted with a simulated unmanned aerial vehicle (UAV) system in a lab-setting to identify and neutralize enemy targets. We recruited and analyzed data from a total of 57 participants (i.e., 15 in high reliability/feedback present group, 14 in high reliability/feedback absent group, 14 in low reliability/feedback present group, and 14 in low reliability/feedback absent group). Contrary to our hypothesis, results indicated that feedback negatively affected users' positive emotions and trust in the system, and increased negative emotions. However, emotions were found to mediate the relationship between feedback and trust, which was expected. Finally, trust was found not to affect the compliance rate with warnings.

In the rest of the paper, we first review prior efforts focusing on trust repair and present our hypotheses. Next, we describe our study design and methodology. Finally, we present our findings along with the implications of our result, followed by limitations of our work to conclude the paper.

PRIOR WORK AND RESEARCH HYPOTHESES

A warning system can fail due to various reasons (e.g., unpredictable execution conditions, nonzero probability of software/hardware failures), and exhibit two types of errors, namely, false alarms and misses [9]. As users often prefer receiving false alarms instead of missing one [46], especially when the cost of a hazard is high, one way system designers attempt to reduce the number of misses is by setting a lower threshold for alarm trigger. However, frequent false alarm is shown to induce the *cry-wolf syndrome*, which can lead to a reduction in subsequent compliance rate with the system [20, 21, 60], and increased response time to signals [21, 31, 60]. While occasional false alarms are often unavoidable, however, they are likely to affect users' trust in the system [49]. This is supported by prior efforts investigating the effect of system failures on trust that confirmed that experiencing even infrequent system errors can lead to distrust in otherwise highly reliable systems [23]. Furthermore, degradation of trust is shown to be immediate after observing system errors [83]. We argue that, as affective process is shown to "ultimately" dominate trust [40], negative emotions triggered due to system errors are likely to play a critical role in degradation of trust. Specifically, system errors are likely to affect *integral affect* which is generated while performing a task, and evolves dynamically with interactions with the system [37, 82]. This is in line with several recent research efforts that acknowledged the influence of affective states on trust [48], and suggested that interaction of user's emotions, moods, and attitudes towards the system define the experience of trust [1, 2, 57, 65].

Prior research looked at various approaches to mitigate the negative effect of automation errors on trust in the system. Among numerous prior efforts focusing on cognitive aspect of trust, Seong et al. demonstrated the importance of providing cognitive feedback regarding system operation in trust calibration [67]. In a more recent work, Wang et al. showed that generating automated explanation regarding a robot's decision-making process helped users to calibrate trust [72]. Along the same line, other recent studies recommended communicating information analysis process [5] and justification behind system's reasoning [55] to maintain an appropriate level of trust. Prior work looked at various approaches for providing feedback as well such as graphical presentation [23, 24], simple textual feedback [7, 13, 18, 62]), and numerical feedback (e.g., confidence score) [67], to name a few.

Another group of work exists that looked at strategies to influence trust through incidental affect (i.e., *incidental affect* refers to the affective state that influences a person's decision-making process of a certain task, even though the source of the affect is unrelated to the task [82]). Various techniques such as showing affective video clips [48], affective images [82], and recalling and writing affective incidents [22] are tried in the past. However, the effectiveness of affect infusion strategies are found to be limited when a user is well familiar with a system/task [48]. Specifically, while incidental affect is shown to have a significant effect on initial trust formation in automated systems, however, over time the effect diminishes as interactions with a system increases [70].

To complement prior efforts, we focus on calibration of trust by influencing *integral affect* through feedback. Specifically, our work is inspired by prior efforts that have shown that delivering positively and negatively framed affective interventions after a system error can improve users' performance significantly (e.g., *"The function of the computer were suspended. Great that the computer will soon work again."*, *"The execution of the program was interrupted. This is frustrating."*) [58]. Furthermore, providing appropriate feedback through affective support system is shown to reduce users' negative affect, and increase positive feelings towards a system, even when the computer itself was the source of the negative affect [36]. Even simple social graces like 'please' and 'thank you' are shown to change users' feelings without being aware of it [59].

Informed by prior efforts, we focus on investigating the interaction effect of feedback and reliability on users' emotions and trust, and develop the following hypotheses to guide our study and analyses.

• ***Hypothesis 1 (H1):*** Giving feedback will result in (a) higher positive emotions, (b) lower negative emotions, (c) higher trust rating, and (d) higher compliance rate.

• ***Hypothesis 2 (H2):*** Higher warning reliability will lead to (a) higher positive emotions, (b) lower negative emotions, (c) higher trust rating, and (d) higher compliance rate.

• ***Hypothesis 3 (H3):*** Higher trust rating will result in higher compliance rate.

• ***Hypothesis 4 (H4):*** Emotions will mediate the relationship between warning reliability, feedback and trust in the system; and (a) positive emotions will lead to a higher trust rating and (b) negative emotions will lead to a lower trust rating.

METHODOLOGY

Design of the Experimental Task

To investigate the aforementioned hypotheses, we designed a 2 (warning reliability: high and low) x 2 (feedback: present and absent) between-group in-lab study where participants

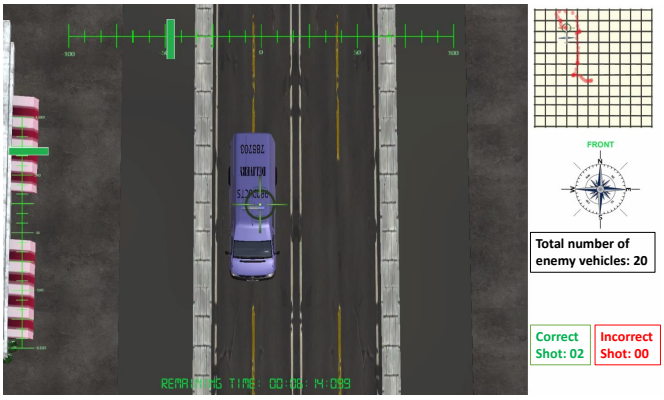

Figure 1: The UAV task screen

completed several rounds of a simulated UAV task to identify and neutralize enemy vehicles in an urban setting.

During the task, participants maneuvered a UAV over a simulated city environment consisting of high-rise buildings and multiple roads. There were a total of 67 vehicles parked on the side of the roads. Vehicles with numbers in addition to text written on top of them were considered enemy vehicles belonging to human traffickers. The other vehicles (i.e., with no text or only text without numbers) were considered innocent vehicles. The objective of the task was to identify and neutralize as many enemy vehicles as possible by shooting them (Figure 1). Participants were explicitly told that there was no human inside the vehicles. All vehicle text consisted of black, non-italic, non-cursive, block fonts that were readable from a distance. We decided to use 20 enemy vehicles per session as we observed that a person could neutralize at most 17 enemy vehicles in the given time during pilot testing the simulation platform. As people tend to perform better when they have a specific goal to accomplish [41, 42], we disclosed the total number of enemy vehicles at the beginning of the task. To further enhance intrinsic motivation, we included the following component regarding social betterment in the description of the task: *"If you successfully neutralize enemy vehicles, you serve your society by making it safer."*

Participants navigated the drone around the city using a Thrustmaster t16000m joystick (Figure 2). They could rotate the drone and change the camera's direction to scrutinize vehicles from different angles. Zoom-in and zoom-out features were provided up to a certain level to ensure that the detection task was not too easy or difficult. The highest zoom-in level allowed users to distinguish numbers from letters easily.

Participants could keep track of the area already explored and plan future directions using a map displayed on the upper right corner of the UAV task screen (Figure 1). A window in the middle right side of the screen was used to display feedback messages accompanied with audio. Below the message window there was a score window displaying the number of correct and incorrect shots taken so far. In the middle bottom part of the screen, a countdown timer displayed the remaining time in the current task session. Each of the four task sessions

Figure 2: Thrustmaster t16000m joystick

was seven minutes long to allow participants develop adequate understanding regarding the system.

Participants were shown their score at the end of each session. In order to make participants careful while shooting, shooting an enemy vehicle was rewarded by a 10 point increase and shooting an innocent vehicle was penalized by a 10 point decrease in the score. At the end of each session, five points were deducted for each enemy vehicle that they failed to neutralize. This was done to create additional time pressure which is shown to affect dependence on automation and trust [61, 71].

Simulated System Error and Warning System

Participants were told that the simulated UAV system had two video transmitters (i.e., a primary transmitter and a secondary transmitter) to transmit videos from the drone to the ground station. The primary transmitter transmitted clear, high-quality video and performed fine most of the time. However, if the primary transmitter got overheated due to continuously transmitting high-quality video, it automatically got restarted and an audio message: "Video will be restored in 30 seconds" was played. It took 30 seconds to cool down and automatically restart the primary transmitter. During this 30 second period, the secondary transmitter took over and transmitted low-quality video. With low-quality video, participants were still able to monitor the city and read texts on vehicles, although it required significantly more effort.

To avoid auto restart of the primary transmitter due to overheating, five seconds prior to the hazard (overheating), an alarm would go off with a warning message ("Transmitter overheating") (one-second duration) followed by a beeping tone for four seconds. Duration of each beep tone (16 bits, 44100 Hz frequency) was 0.18s with an interval of 0.18s between two beeps. Participants could follow the warning by enabling the secondary transmitter before the end of the beeping tone, which would cause low-quality video for 10 seconds and then restore high quality video automatically.

If participants ignored a true positive alarm, the low quality video would start at the end of the warning period (e.g., five sec) and would last for 30 seconds. If participants ignored a false alarm, nothing would happen and the task continued as normal. If participants heeded to an alarm (regardless of true or false alarm), secondary transmitter would engage, and the low quality video would start and last for 10 seconds. This 10 second loss was meant to introduce an element of

	High Reliability	Low Reliability
True Alarm	3	3
False Alarm	1	3
Total Warnings	4	6
Reliability	75%	50%

Table 1: Number of warnings per session across reliability groups.

risk and vulnerability to discourage overreliance on alarms mindlessly [19, 43].

In the current study, we ensured that all the experimental groups experienced three true overheating incidents. Prior efforts suggested that automation reliability below 70% is considered worse than no automation at all [74]. In the current study we followed this general guideline and set the reliability level at 75% (high reliability) and 50% (low reliability) by varying the number of false alarms. Alarm types and numbers are reported in Table 1. All the warnings and hazards occurred in the same order at the same moment for all the participants within the same reliability group. To avoid confounding effect and minimize the number of groups, we only used false-alarm error-bias in the current study. As information regarding system reliability might influence participants' trust in the system [3, 6], we did not disclose the warning reliability to participants.

The Reliability Condition

To prevent participants from considering the system to be 100% reliable, participants were told the following: *"To prevent the primary transmitter from overheating, the drone system has an automated warning module. However, as overheating of the primary video transmitter is related to multiple factors such as the total volume of data transmitted and weather condition, the warning module cannot always analyze the drone system status accurately and thus occasionally provides false-alarms. In case of a true warning, the transmitter will restart after 5 seconds of the warning. So, you will have 5 seconds to respond to a warning. You can follow the warning and enable the secondary transmitter or ignore the warning."* The above was communicated to ensure that participants were aware of the possibility of false alarms and the underlying reasons.

To operationalize the reliability condition, in addition to varying warning accuracy, low-reliability groups experienced a total of 24 alarms over four sessions compared to 16 for high-reliability groups. Our analyses confirmed that, irrespective of the feedback type, low-reliability groups experienced low quality video significantly longer ($M = 69.46$ sec/session) compared to the high-reliability group ($M = 47.80$ sec/session), $F(1, 53) = 71.62$, $p < .01$, $\eta_p^2 = 0.58$. As such, low-reliability condition groups experienced 50% more warnings and significantly longer duration hazy video compared to high-reliability condition groups. Therefore, we concluded that the game playing condition was significantly different between high and low reliability groups.

Note that our design is fundamentally different than having the same number of warnings (e.g., 24) across groups with different true/false ratios, which indeed could have a different effect. Our design was essential to simulate a scenario that would be analogous to many real-life settings where, in the absence of feedback, a user has no way of knowing whether an alarm is true or false. Rather, a user often has to judge the reliability of a system based on the frequency of alarm incidents and the resulting inconvenience, which can influence emotions and trust in the overall system.

Message Designs

Depending on the assigned experimental groups (feedback: absent, present) and response to each warning during the UAV task, participants received one of the messages listed in Table 2. In the feedback present groups, after participants' response to each warning, they received a feedback message informing whether the alarm was true or not with an appropriate affective component (e.g., sorry, thank you).

Alarm Type	User Action	Feedback Absent	Feedback Present
True Alarm	Followed	Video will be restored in 10 seconds.	Thank you for playing it safe! Overheating avoided! Video will be restored in 10 seconds.
	Not Followed	Video will be restored in 30 seconds.	Sorry! Overheating could not be avoided! Video will be restored in 30 seconds.
False Alarm	Followed	Video will be restored in 10 seconds.	Sorry! It was a false alarm. Video will be restored in 10 seconds.
	Not Followed		Sorry! It was a false alarm.

Table 2: Feedback messages based on responses to warnings across feedback groups.

Prior efforts noted that, for effective warning communication, messages should be conspicuous and noticeable [38, 80]. Furthermore, auditory warnings are found to be more noticeable than visual warnings due to the "omni-direction nature" and attention-grabbing capability [80], leading to higher compliance rate compared to non-voice warnings [81]. For that, we used audio messages to make the communication salient, and ensure participants did not miss any message.

Instruments

Personality Traits. As personality traits (e.g., propensity to trust, risk-taking tendencies) is shown to influence decision-making and trusting behaviors when it comes to human-human interaction [17, 47, 69], we measured participants' propensity to trust other people using 12 items rated on a 5-point bipolar Likert scale (i.e., *strongly disagree* to *strongly agree*) [28]. Furthermore, we administered a shortened version of the DOSPERT scale for measuring risk-taking tendencies along five different dimensions: ethical, financial, health/safety, recreational, and social [8, 73].

Emotions. To measure emotions, we adapted three emotion items from four emotion factors proposed by Buck et al.'s UAX (User Affective eXperience) scale, which was originally developed to measure emotions in response to pop-up computer security warning messages [11]. The 12 emotions included: *positive emotions* (i.e., happy, confident, and secure), *anxiety emotions* (i.e., anxious, nervous, and afraid),

loneliness emotions (i.e., lonely, isolated, and abandoned), and *hostility emotions* (i.e., hostile, scornful, and resentful). The questionnaire asked participants to rate emotions in the context of *"How did you feel operating the drone system in your most recent session?"* Each emotion item was phrased in the past tense with a brief statement at the end to put the emotion in context and specifically refer to the emotional experience during the UAV task (e.g., *"I felt happy because the system functioned well."*) Participants rated these statements using a 7-point unipolar Likert scale (i.e., *not at all* to *very much*).

Trust. To measure the trusting beliefs toward the drone system, we adapted the trust scale directly from Chancey et al. [15] which uses the three modified trust factors (i.e., performance, process, and purpose) from the Human-Computer Trust Questionnaire [45]. Here, performance (ability) relates to the observable outcome of an automation tool to achieve a user's goal; process (integrity) reflects the way a automation tool works to advance towards user's goal and understanding of the sequential steps (algorithms) of the automation by a user; and purpose (benevolence) describes why an automation is necessary and a user's knowledge of the automation tool's (and developers') honest intention to maximize user's desired output [15, 39, 40].

In order to avoid possible confusion between the trust in the "overall" system vs. the "warning" system, the items were phrased to measure users' trust in the "overall" system in the paper (e.g., *"Overall, I can rely on the drone system to function properly."*) Sample items measuring performance factor of trust included *"Overall, the drone system performs reliably,"* process factor of trust included *"I recognize how I should use the drone system to perform well the next time I use it,"* and purpose factor of trust included *"If I am not sure about any situation, I have faith that the drone system will operate reliably to help me perform well."* Participants rated their agreement with each item on a 7-point bipolar Likert scale (i.e., *strongly disagree* to *strongly agree*).

Recruitment and Study Protocol

The study was approved by our university's Institutional Review Board (IRB). To recruit participants, we posted recruitment fliers containing a link to the pre-screening survey on notice boards around the university campus. The flier was also sent out to university students, faculty, and staff through an online daily announcement system. We screened potential participants based on the following criteria: (1) normal or corrected-to-normal vision, (2) normal or corrected-to-normal hearing, (3) ability to use both hands to control the joystick, (4) 18 years or older of age, and (5) proficient in English. The chosen criteria ensured that participants would be able understand the feedback messages which were given via a combination of audio and text, and would be able to use both hands at the same time to navigate the UAV during game play.

The experiment was done in an isolated lab environment in the basement of the building where lighting condition and sound were controlled. After arriving at the lab, participants read and signed an informed consent form before starting the experiment. The investigator demonstrated the task and use of the joystick before participants read a detailed description of the task and controls. Next, each participant practiced a 5-minute session in the presence of the investigator to get familiar with the system (e.g., recognize enemy vehicles, use the joystick effectively). Only one participant needed to try the practice session twice. In the practice session, participants did not receive any hazard or warning.

Each participant then responded to the personality traits instruments (i.e., propensity to trust, risk-taking tendencies). Next, they completed a 7-minute UAV task session, which was followed by a survey instrument including emotion items followed by trust items. As we wanted to test whether emotions mediated the relationship between the independent variables and trust, we measured the mediating variable (i.e., emotions) before the outcome variable (i.e., trust) to avoid a possible reverse causal effect [12, 16, 65].

The UAV task sessions were repeated a total of four times. After each task session there was a 5-minute break. At the end of the final session, there was a short interview followed by debriefing. Each in-lab session lasted about 100 minutes and each participant was compensated with a $30 Amazon eGift card. The study took place during April and May of 2018 while participants' data collection time (i.e., morning, noon, or evening) were balanced across groups.

EVALUATION

We evaluated internal reliability of the survey scales (including the extracted emotion factors from UAX scale) using Cronbach's α before running further analyses. A Cronbach's α value greater than 0.8 is considered to have a good reliability [27, 30]. The propensity to trust scale had a good reliability (α = .85). The five risk-taking tendency domains sub-scales did not have good reliabilities (ethical α = .52, financial α = .61, healthy/safety α = .60, recreational α = .77, and social α = .47). Trust in the drone system sub-scales had excellent or good reliabilities (performance factor α = .91, process factor α = .87, and purpose factor α = .89).

We conducted an analysis of a three-way mixed design (split-plot) ANOVA with two between-group factors and one repeated measure factor to test the main effects and interactions for each dependent variable (i.e., emotion factors extracted from Principal Component Analysis (PCA), trust factors, compliance rate, average response time, and UAV task scores). Each particular UAV task session (i.e., of the 4 sessions) was the repeated-measure factor while warning reliability (low, high) and feedback (absent, present) were between-group factors. We inspected normality assumptions, Levene's test of equality of error variances, and consulted appropriate corrective measures when necessary. Our findings along with demographics are presented below.

Demographics

In total, 251 participants completed the pre-screening survey, which included both demographics and screening questions. Out of 239 eligible participants, we recruited 60 participants for the study. After data collection, we removed the responses of one participant who failed to complete the study because of an electrical problem in the lab. In addition, utilizing eye tracker data, we removed the responses of two participants

who completed part of the survey without looking at the survey items. Thus, data from a total of 57 participants were included in our analysis (i.e., 15 in high reliability/feedback present group, 14 in high reliability/feedback absent group, 14 in low reliability/feedback present group, and 14 in low reliability/feedback absent group).

Participants' age ranged from 18 to 27 years (*Mean* = 20.40, *Median* = 20, *SD* = 1.82), while 29 participants were female (50.9%) and 28 were male (49.1%).

The breakdown of reported highest level of education were high school/GED (12.3%; 7), some college degree (66.7%; 38), 2-year college degree (5.3%; 3), 4-year college degree (14%; 8), and master's degree (1.8%, 1). The majority of the participants were undergraduate students (86%; 49), six (10.5%) were graduate students, and two (3.5%) participants were non-students.

Participants' self-reported computer proficiency is as follows: two participants (3.5%) identified themselves as 'beginner,' 30 (52.6%) as 'competent,' 23 (40.4%) as 'proficient,' and two (3.5%) as 'expert.'

Since our study involved operating a simulated UAV with a joystick, we asked participants about their video game playing habits. Only four participants (7%) reported playing video games 'daily,' whereas 15 (26.3%) and 20 (35.1%) participants reported playing a few times a week and a month, respectively. Moreover, 18 participants (31.6%) 'rarely' played video games or 'did not play' at all. We also collected participants experience level with first-person shooting (e.g., Counter Strike, Call of Duty) and air flight combat (e.g., War Thunder, World of Warplanes) video games, where eight participants (14%) reported that they have 'never played' first-person shooting games, 37 participants (64.9%) reported that they 'played before but are not experts', and 12 participants (21.1%) reported themselves as 'expert.' Lastly, 32 participants (56.1%) reported that they have 'never played' any air flight combat game and 25 participants (43.9%) reported that they have, with no participant reporting themselves as 'expert'. We also asked participants about their skill with the joystick used in our study. Reported joystick skill levels were 'never used' (7, 12.3%), 'novice' (30, 52.6%), and 'intermediary' (20, 35.1%). No participants reported being an 'expert' at using the joystick. All participants were right handed. None of the participants served in the military or had experience of operating any military or commercial drone or UAV.

We formed the experimental groups using a randomized complete block design to remove the variability between groups from the experimental error [52]. The aforementioned demographic and game behavior questions were used to balance the composition of experimental groups. We did not find any significant differences across the groups in terms of gender ($\chi^2(3) = .34$, $p = .95$), age ($\chi^2(3) = 3.32$, $p = .35$), highest level of education ($\chi^2(3) = 1.02$, $p = .80$), current employment status ($\chi^2(6) = 4.53$, $p = .61$), computer proficiency ($\chi^2(3) = .10$, $p = .99$), video gaming frequency ($\chi^2(3) = .72$, $p = .87$), first-person shooting game experience ($\chi^2(3) = .37$, $p = .95$), air flight combat games ($\chi^2(3) = 2.06$, $p = .56$), and joystick skills ($\chi^2(6) = 7.76$, $p = .26$). To examine individual differences among the groups, we performed ANOVA analysis with propensity to trust ($F(3,53) = .57$, $p = 0.64$), five domain specific risk-taking tendencies: ethical (χ^2 (3) = 2.77, $p = .43$), financial ($F(3, 53) = .96$, $p = .42$), healthy/safety ($F(3, 53) = .56$, $p = .65$), recreational ($F(3, 53) = .57$, $p = .64$), and social ($F(3, 53) = .63$, $p = .60$).

Based on the aforementioned analyses, we concluded that the groups formed were similar in terms of demographics and personality traits.

Factor Analysis: UAX Emotions Scale

We conducted an exploratory PCA analysis with orthogonal (varimax) rotation to extract factors from the 12 emotion items. For these emotion items, we accumulated the responses of all four sessions together, and performed PCA on the accumulated data to identify the emotion factors. According to Kaiser Criterion, a Kaiser-Meyer-Olkin (KMO) value between 0.8 and 0.9 is "great" [27] and "meritorious" [34], indicating that common factors can explain the variability of the emotion ratings. Our analysis verified KMO measure of sampling adequacy ($KMO = .85$), and Bartlett's test of sphericity measure was significant ($\chi^2(66) = 2005.40$, $p < .001$), indicating that the correlation between items was sufficiently large for PCA. Prior work noted that, for a sample size containing less than 100 participants, all communalities above 0.6 may be "perfectly adequate" [27]. In our study, all communalities were at least 0.73.

	Factor			
	Anxiety	**Positive**	**Loneliness**	**Hostility**
Nervous	**.92**			
Anxious	**.91**			
Afraid	**.89**			
Secure		**.92**		
Happy		**.89**		
Confident		**.88**		
Lonely			**.87**	
Abandoned			**.79**	
Isolated			**.72**	.45
Scornful				**.85**
Resentful				**.80**
Hostile			.55	**.63**

Table 3: Factor loadings of the emotion items from the factor analysis. The highest factor loadings are in bold.

We extracted four factors, which were consistent with Buck et al.'s [11] factor loadings and predicted a cumulative total of 83.33% of the variance in emotion ratings. Factor 1 (anxiety emotions), factor 2 (positive emotions), factor 3 (loneliness emotions), and factor 4 (hostility emotions) explained 44.50%, 21.33%, 12.80%, and 4.70% of the variance, respectively. In addition, the diagonal elements of the anti-image correlation matrix were at least 0.75, which indicated that the factor analysis was appropriate and did not warrant removal of any variable [27]. Furthermore, the positive emotions ($\alpha = .90$), anxiety emotions ($\alpha = .92$), loneliness emotions ($\alpha = .87$), and hostility emotions ($\alpha = .89$) had good reliability. Each factor's subsisting items and their loadings are reported in Table 3.

Main Effects and Interactions on Emotions

Positive Emotions. The ANOVA yielded a significant main effect of feedback on the positive emotions (i.e., happy, confident, and secure) ($F(1, 53) = 5.21$, $p < .05$, $\eta_p^2 = .09$). Feedback absent groups ($M = 4.52$, $SD = .25$) experienced higher level of positive emotions compared to the feedback present groups ($M = 3.72$, $SD = .25$), which was opposite of our expectation (H1a). We did not observe main effect of warning reliability. Hence, our hypothesis H2a was not supported as well.

Anxiety Emotions. We did not observe any main effect or interactions of feedback and reliability on the anxiety emotions (i.e., anxious, nervous, and afraid). As such, our hypotheses H1b and H2b were not supported for anxiety emotions.

Loneliness Emotions. The ANOVA yielded a significant main effect of feedback on the loneliness emotions (i.e., lonely, isolated, and abandoned) ($F(1, 53) = 4.64$, $p < .05$, $\eta_p^2 = .08$). Feedback present groups ($M = 2.03$, $SD = .16$) experienced higher level of loneliness emotions compared to the feedback absent groups ($M = 1.54$, $SD = .17$), which is opposite of our hypothesis H1b for loneliness emotions.

We also observed a significant main effect of reliability on the loneliness emotions ($F(1, 53) = 6.55$, $p < .05$, $\eta_p^2 = .11$). Low reliability warning groups ($M = 2.08$, $SD = .17$) experienced a significantly higher level of loneliness emotions than high reliability warning groups ($M = 1.49$, $SD = .16$), which supported hypothesis H2b for loneliness emotions.

Hostility Emotions. The ANOVA yielded a significant main effect of feedback on the hostility emotions (i.e., hostile, scornful, and resentful) ($F(1, 53) = 8.54$, $p < .01$, $\eta_p^2 = .14$). Feedback present groups ($M = 2.48$, $SD = .19$) experienced higher level of hostility emotions compared to the feedback absent groups ($M = 1.71$, $SD = .19$), which is opposite of hypothesis H1b for hostility emotions.

We also observed a significant main effect of reliability on the hostility emotions ($F(1, 53) = 4.72$, $p < .05$, $\eta_p^2 = .08$). Low reliability warning groups ($M = 2.38$, $SD = .19$) experienced significantly higher level of hostility emotions than high reliability warning groups ($M = 1.81$, $SD = .19$), which supported hypothesis H2b for hostility emotions.

Main Effects and Interactions on Trust Factors

Performance Factor of Trust. We observed a significant main effect of feedback on the performance factor of trust ($F(1, 53) = 5.77$, $p < .05$, $\eta_p^2 = .10$). Regardless of the reliability of warnings, performance ratings were higher for the feedback absent groups ($M = 5.24$, $SD = .24$) than the feedback present groups ($M = 4.45$, $SD = .23$). We did not observe any other main effects or interactions on performance.

Process Factor of Trust. We observed a significant main effect of feedback on the process factor of trust ($F(1, 53) = 5.18$, $p < .05$, $\eta_p^2 = .09$). Regardless of the reliability of warnings, the process factor ratings were higher for feedback absent groups ($M = 5.60$, $SD = .20$) than feedback present groups ($M = 4.95$, $SD = .20$). We did not observe any other main effects or interactions on process.

Purpose Factor of Trust. We observed a significant main effect of feedback on the purpose factor of trust ($F(1, 53) = 4.14$, $p < .05$, $\eta_p^2 = .07$). Regardless of the reliability of warnings, the purpose factor ratings were higher for feedback absent groups ($M = 4.67$, $SD = .26$) than feedback present groups ($M = 3.93$, $SD = .25$). We did not observe any other main effects or interactions on purpose.

We conclude that our findings suggest the opposite of hypotheses H1c and H2c.

Warning Response Behavior

Compliance Rate. Compliance rate is calculated by dividing the number of followed warnings by the total number of warnings. Compliance rates were transformed to a scale of 0 (no compliance) to 1 (full compliance). We did not observe any main effects or interactions on compliance rate. Average compliance rates are reported in Table 4. We conclude that our findings suggest the opposite of hypotheses H1d, H2d, and H3.

Average Response Time. Warning response time is the time a participant took to follow the warning after it went off. We calculated the average response time of each participant by dividing the total warning response time with the number of followed warnings in the task session. We did not observe any main effects or interactions on average response time. Average response times are reported in Table 4.

Warning Reliability	Feedback	Compliance Rate	Average Response Time (sec)	Performance Score
High	Present	.86 (.05)	2.72 (1.98)	109.25 (7.56)
High	Absent	.87 (.05)	2.66 (2.05)	97.86 (7.83)
Low	Present	.83 (.05)	2.62 (2.05)	95.54 (7.83)
Low	Absent	.74 (.05)	2.55 (2.05)	108.93 (7.83)

Table 4: Average compliance rates, response times, and UAV task performance scores. (Numbers outside of parentheses represent means and numbers inside of parentheses represent standard deviations.)

UAV Task Performance Score

We did not observe any significant main effect of warning reliability or feedback on the UAV task performance score. Average performance scores are reported in Table 4.

Mediation Analysis

To establish a mediation, an independent variable must predict the dependent variable [4]. As warning reliability did not predict the trust factors, mediation analysis considering warning reliability as a predictor was not performed. However, we performed simple mediation analyses for the relationship between feedback and trust factors through emotions (e.g., positive, loneliness, and hostility) as this held the required conditions. Mediation analysis results are reported below.

Mediating Effects of Emotions on Performance Factor of Trust. We conducted ordinary least squares (OLS) regression analysis to investigate whether positive emotions mediated the effect of feedback on the performance factor of trust. We used a bootstrap estimation approach with 10,000 samples to test

the indirect effects [25, 68]. As the assumption of homoscedasticity is critical in OLS regression, we used heteroscedasticity-consistent HC3 (Davidson-MacKinnon) standard error estimator [25, 33] to address this assumption. Results indicated that feedback was a significant predictor of positive emotions (b = -.79, SE = .21, $p < .001$) and positive emotions factor was a significant predictor of performance (b = .65, SE = .04, $p < .001$). After controlling for the positive emotions in the full model, feedback was still a significant predictor of the performance factor (b = -.27, SE = .12, $p < .05$), consistent with a partial mediation (Figure 3). Since the 95% confidence interval did not include zero, it indicated that the indirect effect was significant (b = -.52, SE = .14, 95% CI = [-.79, -.25]). Therefore, we conclude that positive emotions partially mediated the relationship between feedback and the performance factor of trust. The regression coefficients (Table 5) indicate that positive emotions increased the performance factor of trust. Hence, the presence of feedback decreased the performance factor of trust by negatively affecting positive emotions.

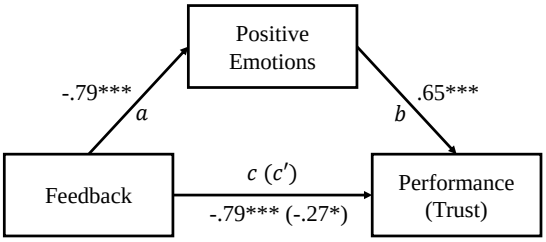

Figure 3: Regression coefficients for a simple mediation model representing the relationship between feedback, positive emotions, and performance factor of trust. The direct effect after controlling for the mediating variable is reported inside the parenthesis. *$p < .05$, **$p < .01$, ***$p < .001$

Source	Positive Emotions	Loneliness Emotions	Hostility Emotions
a	-.79*** (.21)	.48*** (.13)	.77*** (.16)
b	.65*** (.04)	-.44*** (.09)	-.43*** (.07)
c	-.79*** (.18)	-.79*** (.18)	-.79*** (.18)
c′	-.27* (.12)	-.57*** (.17)	-.46** (.17)

Table 5: Regression coefficients for the relationship between feedback and performance factor as mediated by emotion factors. Standard errors are in parentheses. *$p < .05$, **$p < .01$, ***$p < .001$. (Sources a, b, c, and c′ are labeled in Figure 3.)

Simple mediation analysis also indicated a partial mediating effect of loneliness emotions on the relationship between feedback and the performance factor of trust. The indirect effect was significant (b = -.21, SE = .07, 95% CI = [-.35, -.10]). Regression coefficients are reported in Table 5. The regression coefficients indicated that the presence of feedback increased the loneliness emotions, which in turn decreased the performance factor of trust.

Similarly, mediation analysis indicated that the hostility emotions partially mediated the relationship between feedback and the performance factor of trust. The indirect effect was significant (b = -.33, SE = .08, 95% CI = [-.50, -.18]). Reported regression coefficients at Table 5 indicate that the presence of feedback increased hostility emotions, which in turn decreased the performance factor of trust.

Mediating Effects of Emotions on Process Factor of Trust. We observed a partial mediating effect of positive emotions on the relationship between feedback and the process factor of trust (Figure 4). Simple mediation analysis indicated that the indirect effect was significant (b = -.37, SE = .10, 95% CI = [-.58, -.17]). Regression coefficients are reported in Table 6. The regression coefficients indicated that the presence of feedback decreased positive emotions, which in turn decreased the process factor of trust.

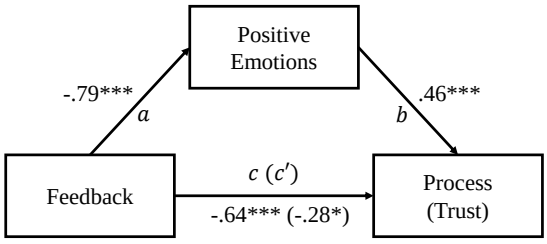

Figure 4: Regression coefficients for a simple mediation model representing the relationship between feedback, positive emotions, and process factor of trust. The direct effect after controlling for the mediating variable is reported inside the parenthesis. *$p < .05$, **$p < .01$, ***$p < .001$

Source	Positive Emotions	Loneliness Emotions	Hostility Emotions
a	-.79*** (.21)	.48*** (.13)	.77*** (.16)
b	.46*** (.04)	-.27** (.09)	-.32*** (.07)
c	-.64*** (.16)	-.64*** (.16)	-.64*** (.16)
c′	-.28* (.13)	-.51** (.15)	-.40** (.15)

Table 6: Regression coefficients for the relationship between feedback and process factor as mediated by emotion factors. Standard errors are in parentheses. *$p < .05$, **$p < .01$, ***$p < .001$. (Sources a, b, c, and c′ are labeled in Figure 4.)

Similarly, mediation analysis indicated that the loneliness emotions partially mediated the relationship between feedback and the process factor of trust. The indirect effect was significant (b = -.13, SE = .05, 95% CI = [-.24, -.05]). Reported regression coefficients in Table 6 indicate that the presence of feedback increased loneliness emotions, which in turn decreased the process factor of trust.

Simple mediation analysis indicated a mediating effect of hostility emotions on the relationship between feedback and the process factor of trust. The indirect effect was significant (b = -.24, SE = .07, 95% CI = [-.39, -.13]). Regression coefficients are reported in Table 6. The regression coefficients indicate that the presence of feedback increased hostility emotions, which in turn decreased the process factor of trust.

Mediating Effects of Emotions on Purpose Factor of Trust. Regression results indicated that the feedback was a significant predictor of positive emotions (b = -.79, SE = .21, $p < .001$) and the positive emotion was a significant predictor of the purpose factor of trust (b = .63, SE = .05, $p < .001$).

After controlling for the positive emotions in the full model, feedback was no longer a significant predictor of the process factor of trust (b = -.23, SE = .15, p = .11), consistent with full mediation (Figure 5). The results indicated that the indirect effect was significant (b = -.50, SE = .14, 95% CI = [-.77, -.24]). Therefore, we conclude that positive emotions fully mediated the relationship between feedback and the purpose factor of trust. The regression coefficients in Table 7 indicate that the presence of feedback decreased positive emotions, which in turn decreased the purpose factor of trust.

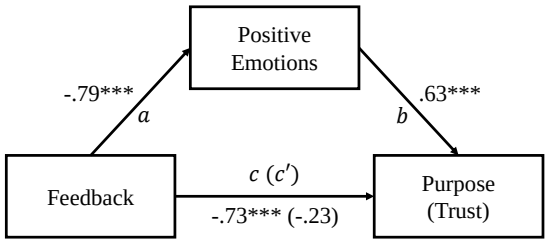

Figure 5: Regression coefficients for a simple mediation model representing the relationship between feedback, positive emotions, and purpose factor of trust. The direct effect after controlling for the mediating variable is reported inside the parenthesis. *$p < .05$, **$p < .01$, ***$p < .001$

Mediation analysis also indicated that the loneliness emotions partially mediated the relationship between feedback and the purpose factor of trust. The indirect effect was significant (b = -.10, SE = .05, 95% CI = [-.20, -.01]). Reported regression coefficients in Table 7 indicate that the presence of feedback increased loneliness emotions, which in turn decreased the purpose factor of trust.

Source	Positive Emotions	Loneliness Emotions	Hostility Emotions
a	-.79*** (.21)	.48*** (.13)	.77*** (.16)
b	.63*** (.05)	-.20* (.09)	-.26*** (.07)
c	-.73*** (.19)	-.73*** (.19)	-.73*** (.19)
c′	-.23 (.15)	-.64** (.19)	-.54** (.19)

Table 7: Regression coefficients for the relationship between feedback and purpose factor as mediated by emotion factors. Standard errors are in parentheses. *$p < .05$, **$p < .01$, ***$p < .001$. (Sources a, b, c, and c′ are labeled in Figure 5.)

Similarly, simple mediation analysis indicated a partial mediating effect of hostility emotions on the relationship between feedback and the purpose factor of trust. The indirect effect was significant (b = -.20, SE = .07, 95% CI = [-.35, -.08]). Regression coefficients are reported in Table 7. The regression coefficients indicate that the presence of feedback increased hostility emotions, which in turn decreased the purpose factor of trust.

Based on our analyses, findings supported hypothesis H4a, whereas hypotheses H4 and H4b received partial support.

DISCUSSION

Negative Effect of Feedback on Emotions and Trust

Contrary to our expectation, feedback present groups reported a lower level of positive emotions compared to the feedback absent groups (opposite of hypothesis H1a). Moreover, results indicated that hostility and loneliness emotions were higher for the feedback present groups. Opposite to our hypothesis H1b, feedback components ("thank you" and "sorry") did not succeed in reducing negative emotions.

Regarding trust, while we found a significant effect of feedback on trust, the effect was opposite of our expectation (Hypothesis H1c). Specifically, participants receiving feedback trusted the system less regardless of warning reliability.

While the effect of feedback was negative on emotions and trust in the system, this might be explained based on prior efforts that noted that, after experiencing a error made by an automated system, users tend to focus more on the error [23, 24]. This is likely due to the fact that users usually have a high reliability expectation from automated systems and expect them to perform reliably with "near perfect" accuracy [24]. For that, even a single error by an autonomous system can cause a significant drop in trust [63, 64]. Therefore, it is possible that explicit feedback messages regarding system's failure to prevent the error (i.e., *Sorry! Overheating could not be avoided!*) affected emotion and trust negatively rather than positively in our case.

While our findings are contrary to prior efforts in human-human interactions that have shown that apology can increase trust [66], however, our findings are in line with other research in the context of unreliable search interface [57] and human-robot interactions [63] that have shown that apologetic messages do not increase trust comparing to neutral messages. Prior research also found that apology without promises of future improvements may be in vain [66], which was true for our message design as well.

We argue that the observed negative effect of feedback on emotion and trust, while unexpected, is not necessarily a "bad" thing. Rather, it might be an effective way to nudge participants to gauge the reliability of automation systems carefully and make decisions while mindfully processing risks. As such, feedback mechanism can facilitate calibration of trust in unreliable systems, preventing the possibility of "overreliance" on automation [56].

Mediating Effect of Emotions on Trust

Regression analysis showed that the positive emotions were positively correlated with the trust factors (i.e., performance, process, and purpose) and the negative emotions (i.e., hostility and loneliness) were negatively correlated with the trust factors. After controlling for the reported experienced emotions, the direct effects of the feedback on trust factors were less dominant or not effective at all. These results support hypothesis H4a and partially support H4b. This result supports the notion that users are likely to consider machines as social actors, and apply social norms of human-human interaction regarding trust to human-machine trust, although most likely subconsciously, which might be exploited to calibrate users' trust in the system. Our findings are in line with prior efforts that have investigated the role of emotions in interpersonal trust, and have shown that positive valance emotional state (e.g., happiness, gratitude) leads to more trust [22]. Influence

of trustor's emotional state on interpersonal trust is supported by others as well [47, 79].

Lack of Effect of Warning Reliability on Compliance Rate

The current study did not find any effect of warning reliability on trust and compliance rate, and hypotheses H2c and H2d were not supported. Several factors may have contributed to this lack of effect. For instance, it is possible that even a single error was enough to reduce trust significantly as users tend to focus more on errors. Another possibility is that, due to the safety-critical nature of the scenario, false alarms did not create a "cry-wolf" effect, and did not significantly affect the compliance rate across groups. This is in line with findings from a previous study in the context of air traffic control conflict alert systems where false alarms did not affect compliance behavior [75]. It is possible that participants might have perceived the risk of not following the warning as higher compared to following (i.e., 30 seconds vs 10 seconds of low-quality video) simulating the characteristics of real safety-critical systems where the cost of noncompliance is high in case of true warning. This cost difference might have caused the high compliance rates regardless of the experimental groups, which satisfies "conditions for dependence" on the system [15]. This finding underscores the difficulty of addressing the danger of "overtrust" in safety-critical autonomous systems (e.g., engaging self-driving feature mindlessly).

Increased time pressure may have been another factor that contributed to this behavior as well. Specifically, in the current study, participants had five seconds to respond to the warnings, and were in imminent threat of experiencing 30 seconds of low-quality video. Furthermore, participants were asked to neutralize as many enemy vehicles as possible within a given time. These can lead to added time pressure, which is shown to increase users' reliance on automation [61, 71]. All these factors might have contributed to the high compliance rate across experimental groups.

Lack of Effect of Trust on Compliance Rate

We did not find any mediating effect of trust on compliance rate, and hypothesis H3 was not supported. This finding is in line with prior efforts that showed that trust measures may not reflect compliance rates [12, 76]. It is likely that trust itself is not the sole mediator of system dependence [16, 40, 78], and should not be considered as the prime mediator of the response behavior [14].

Limitations and Future Work

The study design, feedback messages, and simulation software interfaces were finalized after multiple iterations and pilot testing by the research team that included six members. However, due to the nature of in-lab studies, our simulated scenario has several limitations as follows.

First, we acknowledge that the nature of the simulated task is likely to influence the results, and our findings may not generalize to other domains as we focus on systems with real-time constraints. In fact, our findings are inconsistent with behaviors reported in cybersecurity where users routinely ignore system recommendations [26], underscoring the importance of considering contexts while analyzing such behavior.

Second, as participants might have difficulty identifying emotions and/or might not want to disclose their actual emotions, self-reported emotion ratings may not be perfectly accurate. Further studies in different settings are needed to confirm our findings.

Third, the design of the feedback messages included both gratitude and apology phrasing, which was informed by prior research that demonstrated that apology and appreciation both are necessary components for trust building in human-human interactions [10]. It is possible that excluding these phrasing may cause similar or different effects on emotions, trust, and compliance behavior. Measuring the effect of different feedback messaging could be an interesting follow-up study.

CONCLUSION

In this paper we examined the effect of warning reliability and feedback on users' emotional state, trust, and response behavior in a simulated target detection system. Results indicated that presenting feedback decreased users' trust in the system. In addition, emotions were shown to mediate the relationship between feedback and trust. Our reported findings can be applied to develop executive functioning strategies for safety-critical systems, and design systems that engender appropriate levels of trust instead of simply maximizing it.

ACKNOWLEDGMENTS

This material is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-15-1-0490.

REFERENCES

[1] Yusuf Albayram, Theodore Jensen, Mohammad Maifi Hasan Khan, Ross Buck, and Emil Coman. 2018a. Investigating the Effect of System Reliability, Risk, and Role on Users' Emotions and Attitudes toward a Safety-Critical Drone System. *International Journal of Human–Computer Interaction* (2018), 1–12.

[2] Yusuf Albayram, Mohammad Maifi Hasan Khan, Theodore Jensen, Ross Buck, and Emil Coman. 2018b. The Effects of Risk and Role on Users' Anticipated Emotions in Safety-Critical Systems. In *International Conference on Engineering Psychology and Cognitive Ergonomics*. Springer, 369–388.

[3] Laura H Barg-Walkow and Wendy A Rogers. 2016. The effect of incorrect reliability information on expectations, perceptions, and use of automation. *Human factors* 58, 2 (2016), 242–260.

[4] Reuben M Baron and David A Kenny. 1986. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology* 51, 6 (1986), 1173.

[5] Ellen J Bass, Leigh A Baumgart, and Kathryn Klein Shepley. 2013. The effect of information analysis automation display content on human judgment performance in noisy environments. *Journal of cognitive engineering and decision making* 7, 1 (2013), 49–65.

[6] Matthias Beggiato and Josef F Krems. 2013. The evolution of mental model, trust and acceptance of adaptive cruise control in relation to initial information. *Transportation research part F: traffic psychology and behaviour* 18 (2013), 47–57.

[7] Ann M Bisantz and Younho Seong. 2001. Assessment of operator trust in and utilization of automated decision-aids under different framing conditions. *International Journal of Industrial Ergonomics* 28, 2 (2001), 85–97.

[8] Ann-Renee Blais and Elke U. Weber. 2006. A Domain-Specific Risk-Taking (DOSPERT) Scale for Adult Populations. *Judgment and Decision Making* 1, 1 (2006).

[9] James P Bliss and Richard D Gilson. 1998. Emergency signal failure: Implications and recommendations. *Ergonomics* 41, 1 (1998), 57–72.

[10] Ross Buck. 2014. *Emotion: A biosocial synthesis*. Cambridge University Press.

[11] Ross Buck, Mohammad Khan, Michael Fagan, and Emil Coman. 2018. The User Affective Experience Scale: A Measure of Emotions Anticipated in Response to Pop-Up Computer Warnings. *International Journal of Human–Computer Interaction* 34, 1 (2018), 25–34.

[12] Ernesto A Bustamante. 2009. A reexamination of the mediating effect of trust among alarm systems' characteristics and human compliance and reliance. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 53. Sage Publications Sage CA: Los Angeles, CA, 249–253.

[13] Eric T Chancey, James P Bliss, Molly Liechty, and Alexandra B Proaps. 2015a. False alarms vs. misses: Subjective trust as a mediator between reliability and alarm reaction measures. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 59. SAGE Publications Sage CA: Los Angeles, CA, 647–651.

[14] Eric T Chancey, James P Bliss, Alexandra B Proaps, and Poornima Madhavan. 2015b. The role of trust as a mediator between system characteristics and response behaviors. *Human factors* 57, 6 (2015), 947–958.

[15] Eric T Chancey, James P Bliss, Yusuke Yamani, and Holly AH Handley. 2017. Trust and the compliance–reliance paradigm: The effects of risk, error bias, and reliability on trust and dependence. *Human factors* 59, 3 (2017), 333–345.

[16] Eric T Chancey, Alexandra Proaps, and James P Bliss. 2013. The role of trust as a mediator between signaling system reliability and response behaviors. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 57. SAGE Publications Sage CA: Los Angeles, CA, 285–289.

[17] Jason A Colquitt, Brent A Scott, and Jeffery A LePine. 2007. Trust, trustworthiness, and trust propensity: a meta-analytic test of their unique relationships with risk taking and job performance. *Journal of applied psychology* 92, 4 (2007), 909.

[18] Peter de Vries, Cees Midden, and Don Bouwhuis. 2003. The effects of errors on system trust, self-confidence, and the allocation of control in route planning. *International Journal of Human-Computer Studies* 58, 6 (2003), 719–735.

[19] Munjal Desai, Poornima Kaniarasu, Mikhail Medvedev, Aaron Steinfeld, and Holly Yanco. 2013. Impact of robot failures and feedback on real-time trust. In *Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction*. IEEE Press, 251–258.

[20] Stephen R Dixon and Christopher D Wickens. 2006. Automation reliability in unmanned aerial vehicle control: A reliance-compliance model of automation dependence in high workload. *Human factors* 48, 3 (2006), 474–486.

[21] Stephen R Dixon, Christopher D Wickens, and Jason S McCarley. 2007. On the independence of compliance and reliance: Are automation false alarms worse than misses? *Human factors* 49, 4 (2007), 564–572.

[22] Jennifer R Dunn and Maurice E Schweitzer. 2005. Feeling and believing: the influence of emotion on trust. *Journal of personality and social psychology* 88, 5 (2005), 736.

[23] Mary T Dzindolet, Scott A Peterson, Regina A Pomranky, Linda G Pierce, and Hall P Beck. 2003. The role of trust in automation reliance. *International journal of human-computer studies* 58, 6 (2003), 697–718.

[24] Mary T Dzindolet, Linda G Pierce, Hall P Beck, and Lloyd A Dawe. 2002. The perceived utility of human and automated aids in a visual detection task. *Human Factors* 44, 1 (2002), 79–94.

[25] Andrew F Hayes. 2013. *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach*. The Guilford Press.

[26] Michael Fagan and Mohammad Maifi Hasan Khan. 2016. Why do they do what they do?: A study of what motivates users to (not) follow computer security advice. In *Twelfth symposium on usable privacy and security (SOUPS 2016)*. 59–75.

[27] Andy Field. 2009. *Discovering statistics using SPSS*. Sage publications.

[28] M Lance Frazier, Paul D Johnson, and Stav Fainshmidt. 2013. Development and validation of a propensity to trust scale. *Journal of Trust Research* 3, 2 (2013), 76–97.

[29] Nico H Frijda, Anna Tcherkassof, and others. 1997. Facial expressions as modes of action readiness. *The psychology of facial expression* (1997), 78–102.

[30] Darren George and Paul Mallery. 2001. SPSS for Windows. *Step by Step, A pearson Education Company, USA* (2001).

[31] David J Getty, John A Swets, Ronald M Pickett, and David Gonthier. 1995. System operator response to warnings of danger: A laboratory investigation of the effects of the predictive value of a warning on human response time. *Journal of Experimental Psychology: Applied* 1, 1 (1995), 19.

[32] Michael Grimm, Kristian Kroschel, Helen Harris, Clifford Nass, Björn Schuller, Gerhard Rigoll, and Tobias Moosmayr. 2007. On the necessity and feasibility of detecting a driver's emotional state while driving. In *International Conference on Affective Computing and Intelligent Interaction*. Springer, 126–138.

[33] Andrew F Hayes and Li Cai. 2007. Using heteroskedasticity-consistent standard error estimators in OLS regression: An introduction and software implementation. *Behavior research methods* 39, 4 (2007), 709–722.

[34] Graeme D Hutcheson and Nick Sofroniou. 1999. *The multivariate social scientist: Introductory statistics using generalized linear models*. Sage.

[35] Yung-Tsan Jou, Tzu-Chung Yenn, Chiuhsiang Joe Lin, Wan-Shan Tsai, and Tsung-Ling Hsieh. 2011. The research on extracting the information of human errors in the main control room of nuclear power plants by using Performance Evaluation Matrix. *Safety science* 49, 2 (2011), 236–242.

[36] Jonathan Klein, Youngme Moon, and Rosalind W Picard. 2002. This computer responds to user frustration: Theory, design, and results. *Interacting with computers* 14, 2 (2002), 119–140.

[37] Tamar Kugler, Terry Connolly, and Lisa D Ordóñez. 2012. Emotion, decision, and risk: Betting on gambles versus betting on people. *Journal of Behavioral Decision Making* 25, 2 (2012), 123–134.

[38] Kenneth R Laughery. 2006. Safety communications: warnings. *Applied ergonomics* 37, 4 (2006), 467–478.

[39] John Lee and Neville Moray. 1992. Trust, control strategies and allocation of function in human-machine systems. *Ergonomics* 35, 10 (1992), 1243–1270.

[40] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[41] Edwin A Locke, Dong-Ok Chah, Scott Harrison, and Nancy Lustgarten. 1989. Separating the effects of goal specificity from goal level. *Organizational Behavior and Human Decision Processes* 43, 2 (1989), 270–287.

[42] Edwin A Locke and Gary P Latham. 2002. Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American psychologist* 57, 9 (2002), 705.

[43] Joseph B Lyons and Charlene K Stokes. 2012. Human–human reliance in the context of automation. *Human factors* 54, 1 (2012), 112–121.

[44] Poornima Madhavan, Douglas A Wiegmann, and Frank C Lacson. 2006. Automation failures on tasks easily performed by operators undermine trust in automated aids. *Human factors* 48, 2 (2006), 241–256.

[45] Maria Madsen and Shirley Gregor. 2000. Measuring human-computer trust. In *11th australasian conference on information systems*, Vol. 53. Citeseer, 6–8.

[46] Anthony J Masalonis and Raja Parasuraman. 1999. Trust as a construct for evaluation of automated aids: Past and future theory and research. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 43. SAGE Publications Sage CA: Los Angeles, CA, 184–187.

[47] Roger C Mayer, James H Davis, and F David Schoorman. 1995. An integrative model of organizational trust. *Academy of management review* 20, 3 (1995), 709–734.

[48] Stephanie M Merritt. 2011. Affective processes in human–automation interactions. *Human Factors* 53, 4 (2011), 356–370.

[49] Joachim Meyer. 2001. Effects of warning validity and proximity on responses to warnings. *Human factors* 43, 4 (2001), 563–572.

[50] Joachim Meyer. 2004. Conceptual issues in the study of dynamic hazard warnings. *Human factors* 46, 2 (2004), 196–204.

[51] Joachim Meyer, Chris Miller, Peter Hancock, Ewart J de Visser, and Michael Dorneich. 2016. Politeness in Machine-Human and Human-Human Interaction. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 60. SAGE Publications Sage CA: Los Angeles, CA, 279–283.

[52] Douglas C Montgomery. 2017. *Design and analysis of experiments* (eighth ed.). John wiley & sons.

[53] Clifford Nass and Youngme Moon. 2000. Machines and mindlessness: Social responses to computers. *Journal of social issues* 56, 1 (2000), 81–103.

[54] Clifford Nass, Jonathan Steuer, and Ellen R Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 72–78.

[55] Florian Nothdurft and Wolfgang Minker. 2016. Justification and transparency explanations in dialogue systems to maintain human-computer trust. In *Situated Dialog in Speech-Based Human-Computer Interaction*. Springer, 41–50.

[56] Raja Parasuraman and Victor Riley. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human factors* 39, 2 (1997), 230–253.

[57] S Joon Park, Craig M MacDonald, and Michael Khoo. 2012. Do you care if a computer says sorry?: user experience design through affective messages. In *Proceedings of the designing interactive systems conference*. ACM, 731–740.

[58] Timo Partala and Veikko Surakka. 2004. The effects of affective interventions in human–computer interaction. *Interacting with computers* 16, 2 (2004), 295–309.

[59] Rosalind W Picard and Jonathan Klein. 2002. Computers that recognise and respond to user emotion: theoretical and practical implications. *Interacting with computers* 14, 2 (2002), 141–169.

[60] Stephen Rice. 2009. Examining single-and multiple-process theories of trust in automation. *The Journal of general psychology* 136, 3 (2009), 303–322.

[61] Stephen Rice and David Keller. 2009. Automation reliance under time pressure. *Cognitive Technology* (2009).

[62] Stephen Rice and Jason S McCarley. 2011. Effects of response bias and judgment framing on operator use of an automated aid in a target detection task. *Journal of Experimental Psychology: Applied* 17, 4 (2011), 320.

[63] Paul Robinette, Ayanna M Howard, and Alan R Wagner. 2015. Timing is key for robot trust repair. In *International Conference on Social Robotics*. Springer, 574–583.

[64] Paul Robinette, Ayanna M Howard, and Alan R Wagner. 2017. Effect of Robot Performance on Human–Robot Trust in Time-Critical Situations. *IEEE Transactions on Human-Machine Systems* 47, 4 (2017), 425–436.

[65] Gayle Schwark and Stephen Rice. 2016. An Affect-Trust (AT) Model With Regards to Technological Errors. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 60. SAGE Publications Sage CA: Los Angeles, CA, 1150–1154.

[66] Maurice E Schweitzer, John C Hershey, and Eric T Bradlow. 2006. Promises and lies: Restoring violated trust. *Organizational behavior and human decision processes* 101, 1 (2006), 1–19.

[67] Younho Seong and Ann M Bisantz. 2008. The impact of cognitive feedback on judgment performance and trust with decision aids. *International Journal of Industrial Ergonomics* 38, 7-8 (2008), 608–625.

[68] Patrick E Shrout and Niall Bolger. 2002. Mediation in experimental and nonexperimental studies: new procedures and recommendations. *Psychological methods* 7, 4 (2002), 422–445.

[69] Sim B Sitkin and Amy L Pablo. 1992. Reconceptualizing the determinants of risk behavior. *Academy of management review* 17, 1 (1992), 9–38.

[70] Charlene K Stokes, Joseph B Lyons, Kenneth Littlejohn, Joseph Natarian, Ellen Case, and Nicholas Speranza. 2010. Accounting for the human in cyberspace: Effects of mood on trust in automation. In *Collaborative Technologies and Systems (CTS), 2010 International Symposium on*. IEEE, 180–187.

[71] Casey Tunstall, Stephen Rice, Rian Mehta, Victoria Dunbar, and Korhan Oyman. 2014. Time Pressure Has Limited Benefits for Human-Automation Performance. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 58. SAGE Publications Sage CA: Los Angeles, CA, 1043–1046.

[72] Ning Wang, David V Pynadath, Susan G Hill, and Aberdeen Proving Ground. 2015. Building trust in a human-robot team with automatically generated explanations. In *Proceedings of the Interservice/Industry Training, Simulation and Education Conference (I/ITSEC)*, Vol. 15315. 1–12.

[73] Elke U Weber, Ann-Renee Blais, and Nancy E Betz. 2002. A domain-specific risk-attitude scale: Measuring risk perceptions and risk behaviors. *Journal of behavioral decision making* 15, 4 (2002), 263–290.

[74] Christopher D Wickens and Stephen R Dixon. 2007. The benefits of imperfect diagnostic automation: A synthesis of the literature. *Theoretical Issues in Ergonomics Science* 8, 3 (2007), 201–212.

[75] Christopher D Wickens, Stephen Rice, David Keller, Shaun Hutchins, Jamie Hughes, and Krisstal Clayton. 2009. False alerts in air traffic control conflict alerting system: Is there a "cry wolf" effect? *Human factors* 51, 4 (2009), 446–462.

[76] Rebecca Wiczorek and Dietrich Manzey. 2010. Is Operators' Compliance with Alarm Systems a Product of Rational Consideration?. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 54. SAGE Publications Sage CA: Los Angeles, CA, 1722–1726.

[77] Rebecca Wiczorek and Dietrich Manzey. 2014. Supporting attention allocation in multitask environments: effects of likelihood alarm systems on trust, behavior, and performance. *Human factors* 56, 7 (2014), 1209–1221.

[78] Douglas A Wiegmann, Aaron Rich, and Hui Zhang. 2001. Automated diagnostic aids: The effects of aid reliability on users' trust and reliance. *Theoretical Issues in Ergonomics Science* 2, 4 (2001), 352–367.

[79] Michele Williams. 2007. Building genuine trust through interpersonal emotion management: A threat regulation model of trust and collaboration across boundaries. *Academy of Management Review* 32, 2 (2007), 595–621.

[80] Michael S Wogalter, Vincent C Conzola, and Tonya L Smith-Jackson. 2002. based guidelines for warning design and evaluation. *Applied ergonomics* 33, 3 (2002), 219–230.

[81] Michael S Wogalter, Raheel Rashid, Steven W Clarke, and Michael J Kalsher. 1991. Evaluating the behavioral effectiveness of a multi-modal voice warning sign in a visually cluttered environment. In *Proceedings of the Human Factors Society Annual Meeting*, Vol. 35. SAGE Publications Sage CA: Los Angeles, CA, 718–722.

[82] Jie Xu and Enid Montague. 2015. Affect and Trust in Technology in Teams: The Effect of Incidental Affect and Integral Affect. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 59. SAGE Publications Sage CA: Los Angeles, CA, 205–209.

[83] Kun Yu, Shlomo Berkovsky, Ronnie Taib, Dan Conway, Jianlong Zhou, and Fang Chen. 2017. User trust dynamics: An investigation driven by differences in system performance. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*. ACM, 307–317.

The Apple Does Fall Far from the Tree: User Separation of a System from its Developers in Human-Automation Trust Repair

Theodore Jensen
Department of Computer Science & Engineering
University of Connecticut
Storrs, CT, USA
theodore.jensen@uconn.edu

Yusuf Albayram
Department of Computer Science & Engineering
University of Connecticut
Storrs, CT, USA
yusuf.albayram@uconn.edu

Mohammad Maifi Hasan Khan
Department of Computer Science & Engineering
University of Connecticut
Storrs, CT, USA
maifi.khan@uconn.edu

Md Abdullah Al Fahim
Department of Computer Science & Engineering
University of Connecticut
Storrs, CT, USA
md.fahim@uconn.edu

Ross Buck
Department of Communication
University of Connecticut
Storrs, CT, USA
ross.buck@uconn.edu

Emil Coman
Health Disparities Institute
University of Connecticut
Health Center
Hartford, CT, USA
emil.coman@uchc.edu

ABSTRACT
Responding to automated system errors as violations of user trust can help to promote safe and effective human-computer interactions. Researchers have thus begun investigating mechanisms for "trust repair." However, the extent to which users distinguish between a system and the system's developers is unclear. This may be an important factor in the efficacy of trust repair messages. To investigate this, we conducted a 2 (reliability) x 3 (blame) between-group, factorial study. Participants interacted with a high or low reliability automated system that attributed blame for errors internally ("I was not able..."), pseudo-externally ("The developers were not able..."), or externally ("A third-party algorithm that I used was not able..."). We found that pseudo-external blame and internal blame influenced subjective trust differently, suggesting that the system and its developers represent distinct trustees. We discuss the implications of our findings for the design and study of human-automation trust repair.

Author Keywords
Trust repair; human-automation trust; blame; attribution theory.

CHI'16, May 07–12, 2016, San Jose, CA, USA

ISBN 123-4567-24-567/08/06…$15.00

DOI: http://dx.doi.org/10.475/123_4

CCS Concepts
•**Human-centered computing** → **HCI theory, concepts and models; Empirical studies in HCI;** *Collaborative interaction;*

INTRODUCTION
Trust has been studied in various disciplines as a fundamental factor in human relationships [1, 4, 34]. Additionally, researchers have investigated the notion of a human's "trust" in machines, computers, robots, and automation in general [11, 17]. Given the growing prevalence of automated systems in various domains (e.g., autonomous drone systems, self-driving cars, home assistants), understanding the factors that break and repair a user's trust is increasingly important for system design.

For instance, consider an autonomous drone system used in a surveillance task. If the system misidentifies an image, leaving you exposed to a threat, you may feel that it violated your trust. In a comparatively low-risk scenario, if you use a home assistant system (e.g., Amazon Echo) to manage your calendar and end up missing an important appointment, the experience will likely affect your willingness to rely on the automation in the future. In both cases, the system has the opportunity to repair trust, that is, to respond to the trust violation in an effort to improve future outcomes. We pose the question—*are the people behind an automated system implicated in its mistakes? Or is the system itself deemed a responsible actor that can repair broken trust*?

While users likely correctly understand that these machines are products of human design, evidence suggests that humans respond to computers socially [28]. Perhaps, then, the user is engaging in a trusting relationship the system itself. If so,

users may be prone to poor "trust calibration" [17]. Discounting a fundamental design flaw or lack of functionality as a non-repeatable offense that the system can correct may lead to misuse of the system in unsafe circumstances [29]. On the other hand, diagnosing a rare system malfunction as a serious miscalculation by developers could lead to disuse of the system when it could actually provide an advantage [29]. These represent "overtrust" and "undertrust," respectively.

The current study seeks to elucidate the separation of system and developers by investigating how attribution of blame for system errors influences users' trust. We recruited 147 participants on Amazon Mechanical Turk (MTurk) to play an online game where they collaborated with an Automated Target Detection (ATD) system in 5 rounds of an image classification task. In a 2 (reliability) x 3 (blame) between-group study, participants interacted with a high or low reliability system. After each round, the system displayed a text message acknowledging its errors in identifying images in the previous round, attributing blame either internally ("I was not able..."), pseudo-externally ("The developers were not able..."), or externally ("A third-party algorithm that I used was not able..."). Participants chose how many images to allocate to the automation and were compensated based on their combined performance with the ATD system. After gameplay, participants responded to a survey.

We found that reliability influenced both behavioral and subjective trust, while blame influenced subjective trust. Specifically, internal blame was regarded more positively than pseudo-external blame, suggesting that a system's errors are not considered the same as the developers' errors.

We first review related areas of work in both human-human and human-computer interaction. Then, we describe our study methodology and research hypotheses. Lastly, we present statistical analyses on gameplay and survey data and discuss the implications of our results.

RELATED WORK

In the broad literature on trust, Mayer et al.'s definition is one of the most widely accepted [23]:

> The willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that party.

Despite its original application to trust between humans, this definition readily applies to interactions with automation. For instance, if we trust the Amazon Echo, we expect that it will maintain an accurate, updated schedule of our activities. As we are unable to see inside the black box of the system, our trust involves an acceptance of the risk that it may not end up helping us. If we trust an autonomous vehicle, we expect that it will accurately recognize and respond to obstacles in the driving path. Again, because we cannot fully comprehend the inner workings of the system, our trust exposes us to the vulnerability and potential consequences of an unreliable system. The notion that a non-human may be "trusted" is supported by the Computers are Social Actors (CASA) paradigm.

Computers are Social Actors & Human-Computer Trust

As its namesake suggests, CASA research has found that, in interactions with computers, humans can *"be induced to elicit a wide range of social behaviors, even though users know that the machines do not actually possess feelings"* [28]. For instance, experiments have shown that people use politeness toward computers, as well as social rules regarding praise of others and praise of self (see [28] for more examples). Although research on trust in automation tends to tailor the construct to the trustees' non-human attributes [17, 20], CASA predicts that social rules of trust reserved for other humans are extended to automation.

Yet might such an attribution of trust occur because of the machine's ultimately human creators? Along these lines, "institutional trust" has occasionally been measured as a way to capture a trustor's perceptions of the people behind a system [12, 19, 24]. Similarly, Hoff and Bashir propose that the developers are indirectly the trustees, and that trust in automation may be viewed as *"a specific type of interpersonal trust in which the trustee is one step removed from the truster"* [11]. However, in CASA studies, participants have generally denied responding socially to the computer and remarked that they were not thinking of any human during their interaction. Thus, researchers have suggested that social responses to computers occur due to "mindlessness," as an automatic response wherein individuals *"prematurely commit to overly simplistic scripts drawn in the past"* [27].

Sundar and Nass tested the mindlessness mechanism by investigating whether social responses occur because the computer is perceived as a *medium* between user and developer [37]. In their experiment, one group of participants interacted with a machine consistently referred to as "Computer" while the other group's machine was referred to as "Programmer." Authors predicted that, if the computer were simply considered a medium between user and developer, there would be no difference between experimental conditions. However, those in the "Computer" condition perceived their interaction more positively than those in the "Programmer" condition. Authors suggest that mindlessness prevented participants in the computer condition from considering the developers. Computers appear to be treated as distinct *sources*. We extend this work to the context of trust repair.

Trust Repair

Referring to trust as the "glue" that holds relationships together, Lewicki and Brinsfield note the importance of trust repair following a violation of trust [18]:

> ...it is essential that this glue be rebonded if a broken relationship is to have any hope of returning to a productive or fruitful state.

The authors go on to note various strategies for trust repair that are certainly not uncommon in day-to-day human-human interactions, such as apologies and denials [18].

Recently, de Visser et al. called for human-computer trust researchers to consider systems with the capability of *"building and actively repairing trust"* [7]. In their framework, failures or errors by a system may be viewed as costly *relationship*

acts, while positive interactions such as good performance are considered beneficial relationship acts. *Relationship regulation acts*, which include repair acts and dampening acts, are needed to maintain "*optimal relationship equilibrium*" following costly and beneficial relationship acts, respectively [7]. Relationship regulation by the system can help to maintain a user's appropriately calibrated trust.

In this vein, researchers have studied how users perceive apologies by machines. Tzeng found that, although apologies led to more positive impressions of a computer program in a computer-assisted guessing game, this did not reduce blame for poor game performance [39]. Robinette et al. found that apologies and promises made immediately after a robot's mistake were less effective than the same apologies and promises made at the time users had to make their next reliance decision [33]. Moreover, self-blame has been observed to lead to greater perceived trustworthiness of virtual agents [3].

In an influential study of human-human trust repair, Kim et al. found that the effectiveness of apologies varies with the type of trust violation [15]. Internal apologies (i.e., trustee assumes full responsibility for violation) were more effective than external apologies (i.e., trustee assumes partial responsibility) when the violation was a matter of competence, that is, the trustee's lack of knowledge. However, the opposite was true for an integrity violation, when the trustee knowingly violated trust. Authors suggest that this is a result of the "diagnosticity" of each type of violation—a highly competent individual may demonstrate low competence in some situation, whereas an individual with high integrity is unlikely to demonstrate low integrity. Admitting to low integrity via an internal apology is more diagnostic and, therefore, more hurtful to trust than blaming someone else. Quinn et al. found tentative support to Kim et al.'s findings in the context of human-automation interaction [7, 31].

We build upon these findings and apply attribution theory to observe whether system developers are considered external to the system.

Attribution Theory in Human-Computer Interaction

Attribution theory explores how, in response to a negative event, people attempt to identify and make sense of the event's causes [40]. De Visser et al. suggest that the work of Tomlinson and Mayer [38] on the role of attributional processes in trust repair can be applied to the human-machine context [7].

In Tomlinson and Mayer's model, the locus of causality, controllability, and stability associated with a trust-violating event moderate how perceptions of trustworthiness are affected [38]. *Locus of causality* refers to whether the event was caused by the trustee (i.e., internal) or another actor (i.e., external). *Controllability* is the degree to which an actor had control over the situation. *Stability* reflects the likelihood that the cause will reoccur in the future [40]. Trustworthiness perceptions, also referred to as trusting beliefs, consist of the perceived ability, integrity, and benevolence of the trustee. *Ability* or *competence* consists of the trustee's skills in a particular domain. *Integrity* reflects that the trustee adheres to a set of principles that are acceptable to the trustor. *Benevolence* is the desire of the trustee to do good for the trustor [23].

Social accounts such as apologies and blame can repair trust by managing the trustor's attributions of a trust violation and, in turn, their trustworthiness perceptions [38]. For example, perceptions of the trustee's ability may be less likely to be affected if the trustee convinces the trustor that the event was caused by some external actor or circumstance. We investigate whether developers are considered an external locus of causality following a trust violation by an automated system.

Such a proposition implies that the automation itself acts independently from its developers' control—that computers have agency. In fact, in an interview-based study of 29 computer science majors, Friedman found that 79% of participants judged computers to have decision-making capabilities and 45% judged computers to have intentions, although nearly all participants agreed that a computer's "decision-making" and "intentions" were different from a human's [8]. The study also found that participants who did not blame a computer for errors gave reasons that diminished its agency. Those who did blame the computer often reasoned about its participation in the events leading up to the error.

Researchers have further investigated this idea of computers as "scapegoats" to which blame can be attributed. For example, a self-serving bias has been identified where people are less willing to blame computers that are similar to themselves [26]. Moreover, the greater degree of autonomy of an agent [36] as well as mere perceptions of a teammate as human or AI [25] have been observed to influence the assignment of blame. Prior work has found that minor linguistic manipulations in a virtual driving assistant's messages (e.g., using "you" instead of "we") can influence the degree to which a driver attributes responsibility to the system, as well as their perceptions of the system in general [14].

We build on this work by manipulating the locus of causality for system errors in trust repair messages, in order to observe what happens when the developers are a target of blame.

METHODOLOGY

To investigate the extent to which users distinguish between system and developers, as well as how this varies with system reliability, we designed a 2 (reliability: high, low) x 3 (blame: internal, pseudo-external, external) between-group, factorial study with the following hypotheses.

First, we expected a main effect of reliability on participants' trust:

> H_1: High reliability leads to greater trust in the system than low reliability.

Moreover, in line with CASA, we anticipated that blame of the system would have a different effect than blame of the developers:

> H_2: Pseudo-external blame (i.e., blame of developers) will influence trust differently than internal blame (i.e., blame of the system itself).

Internal	"I am sorry that X images assigned to me were counted as misidentifications. I was not able to process those images."
Pseudo-External	"I am sorry that X images assigned to me were counted as misidentifications. The developers were not able to account for processing those images."
External	"I am sorry that X images assigned to me were counted as misidentifications. A third-party algorithm that I used was not able to process those images."

Table 1. *Feedback messages*. Based on a participant's *blame* condition, these messages were identical across *reliability* groups. "X" represents the number of images that the automation was unable to identify in the previous round.

We designed an online game where participants were scored and compensated based on their performance with an automated system of high or low reliability. After each round, the system displayed a feedback message acknowledging errors in the previous round. Feedback messages in each blame condition are shown in Table 1. After gameplay, participants responded to survey on their perceptions of the system. The details of our methods follow.

Game Design

In the "Target Identification Task," participants had to classify 20 images of vehicles as "Dangerous" or "Not Dangerous" in each of 5, two-minute rounds with the help of the Automated Target Detection (ATD) system.

The game represented a task where two drones (one automated, one manually controlled) monitor a large area for criminal activity. After clicking on a map marker, the manual drone icon moved toward that location. Upon arriving, an image of a vehicle was shown in the "Vehicle Identification Panel." Participants then used "Zoom In," "Zoom Out," and "Rotate" buttons to manually inspect n images, where $0 \leq n \leq 20$. Non-dangerous vehicles had only text on top of them. Dangerous vehicles had numbers in addition to text. Correct and incorrect manual identifications were accompanied by a bell and buzzer sound, respectively. A timer counting down from 2 minutes was shown in the upper right part of the game interface, which is shown in Figure 1.

The automation ostensibly worked in parallel on (20 - n) images (it did not actually process images, per se—it spent a fixed amount of time and had accuracy determined by the reliability condition). In the first round, n was set to 10. In later rounds, participants were able to choose how many images to allocate to the automation beforehand. This measure was used to characterize behavioral trust, or reliance on the automation.

A similar drone-based monitoring game was used by Satterfield et al. to study trust [35]. Our design differs in that the participant and automation control only one UAV each, rather than multiple assets. Moreover, rather than intervening during gameplay, our participants allocate control to the automation before a given round. Lastly, while Satterfield et al. give participants extensive practice before their single experimental session, we decided not to use a practice round in order to observe participants' trust development as they grew familiar with the system.

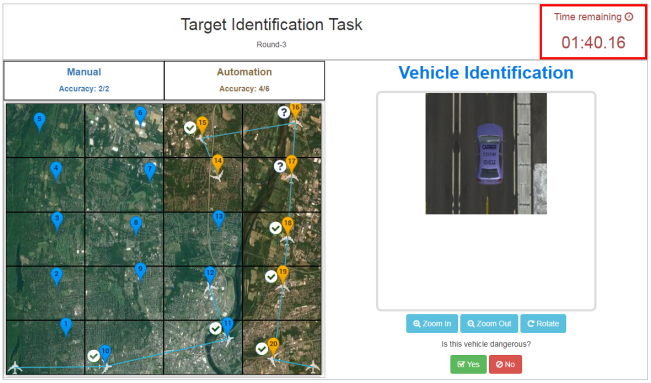

Figure 1. *Target Identification Task interface*. On the left, the map with a marker for each image is shown. A check mark is displayed at locations of correctly classified images, an "X" for incorrectly identified manual images, and a "?" for automation images that could not be processed. Accuracy for both parties is updated in real-time above the map. On the right, the Vehicle Identification Panel shows the current manual image. Participants answer "Is this vehicle dangerous?" with the "Yes" or "No" button after using the "Zoom In," "Zoom Out," and "Rotate" buttons to inspect the image. The timer at the upper right counts down from 2 minutes.

Scoring

For each round, participants were given a 100-point "Round Score" crediting their collective speed and accuracy with the automation. Speed was ostensibly important in order to identify criminals before they could harm people; accuracy so that resources were not inefficiently devoted to stopping innocent people. The Round Score was calculated by averaging a participant's "Time Score" and "Accuracy Score." The Time Score credited those with more time remaining in the 2-minute round. The Accuracy Score was based on the correctness of the user's and the automation's combined performance. The number of correct identifications, $n_{correct}$, for the automation in a given round was calculated as follows:

$$n_{correct} = floor(r * A_k)$$

where r was 0.6 and 0.9 for low and high reliability groups, respectively, and A_k represents the number of images allocated to the automation in the k^{th} round. For example, if 13 images were allocated to the low reliability automation before a round, there would be 7 correct identifications and 6 images unable to be processed by the automation (i.e., "counted as misidentifications"). As a result, the automation improved speed, but could compromise accuracy. We pilot tested various scaling functions for the Time and Accuracy scores in order to produce Round Scores that discouraged full allocation to the automation or fully manual identification. The reliability condition thus allowed us to observe how participants calibrated their trust to an appropriate level as they grew familiar with the system's capabilities [17]. The accuracy in each reliability condition was chosen based on prior work identifying 70% as the level at which a system is considered reliable [41].

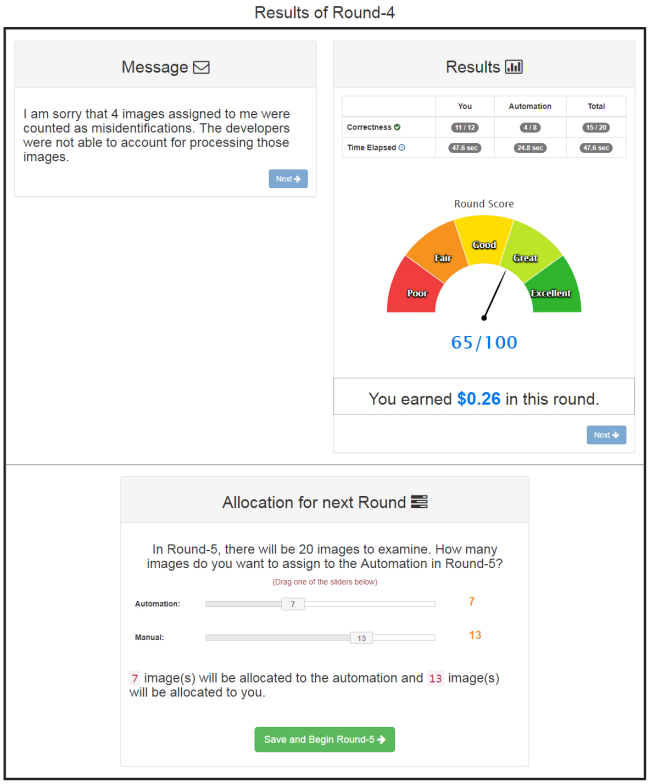

Figure 2. ***Feedback page*****. The feedback message is shown at the upper left. The upper right scoring panel contains a table with the correctness and time elapsed for manual and automated identification. The "Total" column shows the longer of the two times and combined correctness, which contribute to Time Score and Accuracy Score, respectively. The Round Score is shown with a colored gauge and its associated bonus compensation amount (see Compensation subsection in Methodology). In the lower panel, participants choose how many images to allocate to the automation for the next round. Adjusting either the manual or automation slider simultaneously adjusts both values.**

In order to motivate good performance and appropriate trust calibration in the game, we compensated participants for their cumulative Round Scores at a rate of 100 points = $0.40.

Feedback Page

After each round, participants were shown a feedback page containing three elements shown in Figure 2: 1) feedback message, 2) score information, and 3) allocation decision. The feedback messages for each group are shown in Table 1. All messages were displayed with a typewriter effect. Raw scoring information about the time and accuracy of both manual and automated identification was next displayed in a table, underneath which a participant's Round Score was shown. A colored gauge indicated where the score fell out of a 100 possible points along with its associated compensation amount. Lastly, participants were asked to choose how many images to allocate to the automation for the following round using a slider. A dialog box asked for confirmation of this choice before the next round of gameplay was started. Each panel had a time-delayed "Next" button to prevent participants from quickly advancing through the page.

Post-Gameplay Survey

The survey was hosted on our university's Qualtrics server. To measure perceptions of the automation in the game, as opposed to the overall game interface or the computer they were using to complete the study, we explicitly noted that the phrase "the system" throughout the survey would refer to the ATD system that helped participants identify images during gameplay. Participants first were required to correctly answer a multiple choice question ensuring this understanding before they advanced to the survey.

The survey began with demographic questions on gender, age, race, education level, experience operating drones or UAV's, military service, and frequency of video game playing. Manipulation check questions were also included to verify the effect of our independent variables and framing of the messages.

We next investigated participants' attributions of responsibility for their performance in the game using 2 items adapted from Moon and Nass [26]. Participants used a 10-point slider from "You" to "Automation" indicating who was more responsible for 1) overall performance in the game and 2) Round Scores.

To complement our behavioral measure of trust, we measured trusting perceptions, or subjective trust, in the survey. Researchers have measured trust in automation in various ways. Jian et al.'s scale [13] is one of the most widely cited and has been applied to the study of apology messages [30]. In general, such instruments [5,21] incorporate dimensions similar to Mayer et al.'s trustworthiness characteristics: ability, integrity, and benevolence [23]. For example, Madsen and Gregor's scale consists of five constructs that reflect these elements: perceived reliability, perceived technical competence, perceived understandability, faith, and personal attachment [21]. Researchers investigating trust in e-Commerce [24] and online recommendation agents [2] have directly adapted the trusting beliefs from Mayer and Davis [22]. However, such an approach has not been taken in the trust in automation literature. For this reason, we measured perceptions of trust in the ATD system with 1) trusting belief items from McKnight et al. [24] (originally from Mayer and Davis [22]) and 2) Jian et al.'s trust in automation scale [13], both rated on 7-point scales. Subjective trust items are included in the Appendix.

We also included 4 attention check questions throughout the survey.

Study Procedure

The study was posted as a Human Intelligence Task (HIT) on MTurk, available to workers 18 years or older, living in the United States, and having completed at least 1000 HIT's with an approval rating of 95%. When participants accepted the job on MTurk, a link directed them to the game website. The first page of the game displayed an information sheet, the end of which asked participants whether or not they consented for participation in the study.

Participants who gave consent were then brought to an instruction page explaining the task that the game represented, the importance of speed and accuracy in scoring, and game controls. At the end of this page, participants were required

to correctly answer a series of questions to ensure their understanding of these aspects. One question ensured they were using audio to hear the sound effects. If a participant incorrectly answered a question, they remained on the instruction page until they answered all correctly. No feedback was given as to which answer was incorrect.

Participants were then shown the following introduction message: "Welcome! I will use these messages to communicate with you about my performance." After advancing, they began the first round of the game. We told participants beforehand that certain behaviors would prevent them from completing the study. These included inactivity for a round where there were manual images to identify, refreshing the webpage, or clicking back to return to the previous webpage.

Data collected from gameplay were anonymous and linked to survey responses with a participant's randomized ID. Those who completed all rounds of gameplay and the survey submitted their random, Qualtrics-generated ID into MTurk to be compensated with $2 for completion of the HIT. We utilized MTurk's bonus payment feature for the aforementioned bonus compensation based on cumulative Round Scores. The study was approved by our university's IRB.

EVALUATION

A total of 264 participants completed the study. We first conducted a series of screening procedures to ensure that our sample consisted of attentive MTurk users. We removed data for 13 participants who incorrectly answered at least 1 attention check question in the survey. Additionally, we removed data of 7 participants who allocated all 20 images to the automation in each of the final 4 rounds, as this strategy indicated a clear lack of motivation and regard for the scoring mechanism. There were no participants who used fully manual identification in the final 4 rounds. For the remaining participants, we next looked to our manipulation check questions.

Manipulation Checks

Because our focus was on the locus of blame in the messages, we wanted to ensure that participants recognized 1) that the source of the messages was the computer itself, and 2) the target of blame that aligned with their group.

We first asked participants what entity was communicating with the messages, with options "The computer," "The system developers," "Not sure," and "Other." Since all of those who chose "Other" mentioned either the "the system," "the automation," or "ATD," they were coded as correct answers along with those who answered "The computer."

Next, we asked what entity was mentioned in the message as the cause of system errors, with options "The computer," "The system developers," "A third-party algorithm," "Not sure," and "Other." For those who chose "Other," who all were in an internal blame condition, if the participant mentioned that it was the system that was unable to process images, these were coded as correct answers.

Initial group sizes, the number of participants in each group who failed source and blame manipulation checks, and group sizes after screening out incorrect answers are shown in Table 2. The high rate of source failure in pseudo-external conditions may have resulted from explicit mention of the system developers. This may have prompted participants to think about the developers' role in creating the system's messages despite the use of "I." Likewise, the high rate of blame failure in internal blame conditions may have resulted from the target of blame being given implicitly (i.e., "I was not able to...") rather than explicitly as in the other conditions (e.g., "The developers were not able to...").

	$n_{initial}$	Fail. Source	Fail. Blame	n_{final}
Low-Int.	38	4 (10.5%)	17 (44.7%)	**21**
Low-Psuedo.	42	12 (28.6%)	13 (31.0%)	**19**
Low-Ext.	43	10 (23.3%)	5 (11.6%)	**29**
High-Int.	41	6 (14.6%)	14 (34.1%)	**25**
High-Pseudo.	40	16 (40.0%)	9 (22.5%)	**22**
High-Ext.	40	7 (17.5%)	2 (5.0%)	**31**

Table 2. *Group sizes and manipulation check failure rate*. The percentages of participants within each group who *failed* the source and blame manipulation checks are shown. Some participants failed both manipulation checks.

All subsequent analyses are conducted on the remaining 147 participants, with group sizes shown as n_{final} in the last column of Table 2.

To test the efficacy of our reliability manipulation, we asked participants to report how many images they would expect the automation to correctly identify out of 100 using a slider. Those in both the low (Mean = 56.1, Median = 60.0, SD = 11.1) and high (Mean = 82.7, Median = 85.0, SD = 14.4) reliability conditions accurately assessed the automation's reliability[1]. A Mann-Whitney U-test confirmed that the difference in perceived reliability between groups was significant (U = 5,164.00, p <0.001) and that our reliability manipulation was effective.

Sample Demographics

Of the 147 remaining participants, 83 (56.5%) were male and 64 (43.5%) female. The average age was 34.9 years (Median = 33.0, SD = 9.5). There were 115 (78.2%) white/Caucasian, 12 (8.2%) African American, 10 (6.8%) Asian, 7 (4.8%) Hispanic, and 3 other participants. Of these, 85 (57.8%) reported having at least a 4-year college degree, 144 (98.0%) having never operated a military or commercial drone or UAV, and 10 (6.8%) having served in the military. Lastly, 57 (38.8%) participants reported playing games on their computer or mobile device daily, 57 (38.8%) a few times a week, and the remaining 33 (22.4%) a few times a month or less.

A Chi-Square test revealed that there were no significant differences between groups in terms of gender ($\chi^2(5) = 2.78$, p = 0.73). Using Fisher Exact Tests to handle cells with less than 5 participants, we found no significant differences between groups in terms of drone or UAV experience (p = 0.43), military service (p = 0.75), or race (p = 0.48). Moreover, using

[1]While the mean of high reliability ratings is slightly below 90, this number is relatively accurate since allocating 11 or more images to the automation corresponded to 2 errors.

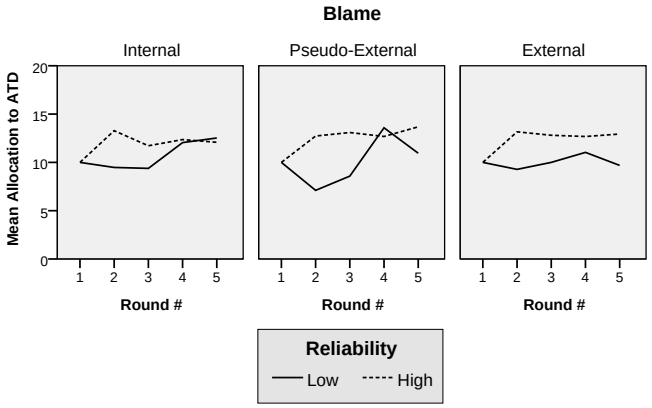

Figure 3. ***Group mean allocation to ATD*****. Means for each group's allocation to automation over the 5 rounds of gameplay are shown. Error bars are excluded for ease of viewing.**

Kruskal-Wallis Tests, we found no significant differences in terms of age ($\chi^2(5) = 3.31$, $p = 0.65$), education level ($\chi^2(5) = 5.23$, $p = 0.39$), or gaming frequency ($\chi^2(5) = 7.41$, $p = 0.19$). We concluded that the groups are demographically similar and can be compared based on our manipulations.

The instruction page and 5 rounds of gameplay took an average of 11.1 minutes (Median = 10.2, SD = 4.2) while the survey took 9.8 minutes (Median = 9.3, SD = 3.9)[2].

Gameplay Behavior

For each group, the average number of images allocated to automation in each of the 5 rounds of gameplay is shown in Figure 3. The first, second, and third plots show internal, pseudo-external, and external blame conditions, respectively. Low and high reliability groups are shown as separate lines in each plot.

In general, Figure 3 gives an idea of how participants in each group calibrated their trust in the ATD system over time. The effects of reliability are immediately clear.

Blame conditions also appear to have influenced participants' willingness to allocate to the automation, especially entering Round 2 after initial exposure to the feedback message. For example, in the low reliability condition, pseudo-external participants reduced their allocation by an average of nearly 3 images (M = -2.89, SD = 2.33) prior to Round 2 while internal (M = -0.52, SD = 3.64) and external (M = -0.72, SD = 3.70) participants reduced their allocation by an average of less than 1 image. These results should be interpreted with caution given the large standard deviations of the metrics.

Analysis of Variance

To investigate our hypotheses, we conducted a two-way, 2 (reliability) x 3 (blame) Analysis of Variance (ANOVA) for various dependent variables.

[2]Reported survey time accounts for other items that are not mentioned in this paper due to space constraints.

First, we combined the 2 attributional items for overall performance and Round Scores into a single Attribution of Responsibility measure because they were highly correlated (Spearman's $\rho = 0.79$, $p < 0.001$). Higher values indicate that a participant considered the automation as more responsible than them for outcomes in the game.

Next, behavioral trust was operationalized as Total Allocation (TA) and First Calibration (FC):

$$TA = \sum_{k=1}^{5} A_k$$

$$FC = A_2 - 10$$

where A_k indicates the number of images allocated to the automation for the k^{th} round of gameplay. TA gives a sense of overall trust in the system and FC gives an impression of participants' immediate reactions to Round 1 gameplay and feedback messages. FC values are negative for participants who allocated fewer than 10 images for Round 2 and positive for those who increased their allocation above 10.

Lastly, subjective trust measures consisted of Jian et al.'s trust in automation scale (12 items, Cronbach's $\alpha = 0.88$) and the perceived trustworthiness characteristics: ability (4 items, $\alpha = 0.95$), integrity (4 items, $\alpha = 0.88$), and benevolence (3 items, $\alpha = 0.71$). All of the subjective trust scales demonstrated acceptable reliability [9].

The correlations between dependent variables are shown in Table 3. All post-hoc pairwise comparisons between blame conditions were done using Tukey's HSD. We report effect size as partial eta-squared, η_p^2, which represents the proportion of variance explained by a predictor relative to the error term in a given model [6].

Since the attributional measure was the only dependent variable to yield an insignificant ANOVA model, we focus on trust measures. The details of our analyses follow.

Total Allocation

There was a significant main effect of reliability on Total Allocation ($F(1,141) = 30.36$, $p < 0.001$, $\eta_p^2 = 0.177$). Participants in the high reliability condition (M = 51.06, SD = 9.28) allocated more images to the automation overall than those in the low reliability condition (M = 41.10, SD = 11.95).

First Calibration

There was also a significant main effect of reliability on First Calibration ($F(1,141) = 57.86$, $p < 0.001$, $\eta_p^2 = 0.291$). While participants in the high reliability condition (M = 3.08, SD = 3.54) increased their allocation to automation following Round 1, those in the low reliability condition (M = -1.26, SD = 3.47) decreased their allocation.

Trust in Automation

The ANOVA on Trust in Automation revealed a significant main effect of reliability, ($F(1,141) = 45.82$, $p < 0.001$, $\eta_p^2 = 0.245$). High reliability participants (M = 5.16, SD = 0.89) trusted the system more than low reliability participants (M = 4.13, SD = 0.86).

	1	2	3	4	5	6
1– Total Allocation	–					
2– First Calibration	0.71**	–				
3– Trust in Automation	0.45**	0.47**	–			
4– Ability	0.38**	0.47**	0.79**	–		
5– Integrity	0.09	0.18*	0.50**	0.55**	–	
6– Benevolence	0.29**	0.29**	0.51**	0.53**	0.65**	–
7– Attribution of Responsibility	0.44**	0.18*	0.11	0.16*	0.02	0.20*

Table 3. ***Correlation between dependent variables*****. Spearman's ρ between each pair of variables are shown.** $^{**}p < 0.01$**,** $^{*}p < 0.05$**.**

Next, we conducted a Multivariate Analysis of Variance (MANOVA) for the perceived trustworthiness characteristics. As this test revealed significant main effects of reliability ($F(3,139) = 24.29$, p <0.001, Wilks' $\lambda = 0.656$, $\eta_p^2 = 0.344$) and blame ($F(6,278) = 4.01$, $p = 0.001$, Wilks' $\lambda = 0.847$, η_p^2 = 0.080), we conducted an ANOVA for each characteristic, using a Bonferroni-adjusted significance level of $\alpha = 0.05/3 = 0.0167$.

Ability

There was a significant main effect of reliability ($F(1,141) = 65.89$, p <0.001, $\eta_p^2 = 0.318$)[3]. High reliability participants (M = 5.20, SD = 1.03) had significantly greater perceptions of the system's ability than low reliability participants (M = 3.53, SD = 1.35).

Integrity

There were significant main effects of reliability ($F(1,141) = 7.05$, $p = 0.009$, $\eta_p^2 = 0.048$) and blame ($F(2,141) = 9.97$, p <0.001, $\eta_p^2 = 0.124$) on integrity. High reliability participants (M = 4.88, SD = 1.18) had significantly greater perceptions of the system's integrity than low reliability participants (M = 4.30, SD = 1.23). Moreover, post-hoc analysis revealed that internal blame participants (M = 5.17, SD = 1.20) had significantly greater perceptions of integrity than both pseudo-external (M = 4.07, SD = 1.03) (p <0.001) and external participants (M = 4.55, SD = 1.23) ($p = 0.016$).

Benevolence

There were significant main effects of reliability ($F(1,141) = 15.58$, p <0.001, $\eta_p^2 = 0.100$) and blame ($F(2,141) = 5.68$, p = 0.004, $\eta_p^2 = 0.075$) on benevolence. High reliability participants (M = 4.51, SD = 1.14) had significantly greater perceptions of the system's benevolence than low reliability participants (M = 3.69, SD = 1.25). Additionally, internal blame participants (M = 4.57, SD = 1.17) had perceptions of benevolence that were significantly greater than pseudo-external participants (M = 3.75, SD = 1.12) ($p = 0.003$), and marginally significantly greater than external participants (M = 4.04, SD = 1.31) ($p = 0.050$).

[3]Levene's test revealed that the variance in ability beliefs across groups was not equal. However, the p-value for this effect was substantially below our threshold, suggesting that the result is robust in the face of heterogeneous variances.

DISCUSSION

Overall, supporting H_1, we found a consistent main effect of reliability, wherein high reliability participants demonstrated greater behavioral trust and reported greater subjective trust than low reliability participants. Also, although blame did not significantly affect behavior, we found partial support for H_2 in that pseudo-external blame caused lower perceptions of the system's integrity and benevolence than internal blame, suggesting that the system's mistakes were not inherently attributed to its developers. Instead, it appears that framing errors as the developers' responsibility led to more negative perceptions of the system's trustworthiness.

We discuss this apparent separation of system and developers in the context of the Target Identification Task, as well as the implications of this finding and possible directions for future trust repair research.

The Target Identification Task as an Environment of Human-Automation Collaboration and Trust

The online game we developed appears a ripe context for the study of human-automation trust. Applying our operational definition from Mayer et al. [23], participants were not able to "control" the ATD system during the game. They could only take a leap of faith before each round in deciding how many images to allocate to it. This represented a "willingness... to be vulnerable to the actions" of the system, with the "expectation" that the system would assist in achieving good performance.

To create the perception of vulnerability and risk that is necessary for trust, the scoring function in the game penalized lower accuracy that could result from using the automation. To create a need for reliance on the automation, it penalized slow speed that could result from manual identification. The main effect of reliability lends to the effectiveness of this scoring mechanism. In support of H_1, high reliability participants chose to allocate more images to the automation than low reliability participants over the course of gameplay. This effect carried over into subjective reports of the system's ability, integrity, and benevolence.

Contextual factors of this particular study, such as MTurk users' incentive to quickly complete HIT's, may influence the validity of allocation as a behavioral measure of trust. While we controlled for this with bonus compensation and attention checks, and corroborated with subjective trust measures, in-lab studies may reduce this particular bias of an MTurk sample.

We believe our findings generalize to other instances of human collaboration with or use of automated systems, yet recognize that some aspects of the game may not apply to practical systems. For instance, in order to quantify behavioral trust, we allowed participants a spectrum of control over the automation. In practical systems, reliance may be a binary decision where a user decides whether or not to use automation. We encourage studies that build on our findings in contexts where reliance takes the latter form. Also, the ATD system performed with consist accuracy in both reliability conditions. As systems become increasingly autonomous, future studies should consider what type of relationship regulation acts can maintain appropriately calibrated trust if system behavior and reliability are dynamic.

Conceptual Separation of Developers and the System

The crux of our experiment lies in that internal and pseudo-external blame have the same technical implications for the system. The messages "I was unable to process those images" and "The developers were not able to account for processing those images" should theoretically have the same effect on trust, since the system's inability is ultimately a result of its developers not being able to give it some functionality. Despite this, we found support for H_2 in that there were differences in trusting perceptions between internal and pseudo-external blame conditions.

Specifically, internal blame led to significantly greater integrity perceptions than pseudo-external and external blame. The difference between the latter two groups was not significant. Internal blame also led to significantly greater benevolence perceptions than psuedo-external blame. Although we did not find significant results in terms of behavioral trust, group means for First Calibration indicate that these reduced integrity and benevolence perceptions may have initially impacted allocation decisions. Understanding the situations in which perceived integrity and benevolence influence reliance behaviors is a promising area of future research.

These findings coincide with those where self-blame was associated with more positive perceptions of a robot [10] and virtual agent [3] than blame of other parties. Yet further, the fact that mentioning developers led to differences in perception suggests that developers are actually considered an *other* relative to the system. This is in line with Sundar and Nass' finding that computers are distinct sources and not mere media between user and developer [37], and suggests that trust in an automated system is not exactly the same as trust in its developers. In other words, errors that are attributed to developers appear to influence trust differently than errors attributed to the system itself. This manifested in perceptions of the system's trustworthiness in our study.

The Ability, Integrity, and Benevolence of Non-Human Trustees

The trust in automation literature generally conceptualizes and measures subjective trust in ways specifically crafted for such non-human trustees. For example, Lee and Moray [16] suggested performance, process, and purpose as bases for trust. These align roughly to Mayer et al.'s ability, integrity, and benevolence [23]. Taking a CASA approach, we directly applied Mayer et al.'s human-human trustworthiness characteristics to the automated system in our study.

It is important to note that each trusting belief item referred directly to "the system" as the trustee. We intentionally did not mention the developers. While some items may have appeared illogical (e.g., "The system is concerned about my well-being, not just its own," "The system is sincere and genuine"), we found that responses in each group were approximately normally distributed and moderately correlated with Jian et al.'s trust scale. It appears that participants were able to conceptualize the automated system's concern and sincerity.

In general, the trusting beliefs seem to paint a more thorough picture of the trustworthiness of automated systems. Not only did low reliability predictably reduce perceptions of ability, integrity, and benevolence, but manipulations in message content influenced integrity and benevolence perceptions. In fact, the effect of blame on perceived integrity ($\eta_p^2 = 0.124$) was nearly three times the size of the effect of reliability ($\eta_p^2 = 0.048$). This effect of blame on perceptions of the system was not captured by Jian et al.'s scale.

In some sense, blame of developers may have reduced integrity perceptions because it seemed particularly hypocritical on the part of the system (integrity items refer to the system being "truthful" and "genuine"). Yet, this may also provide evidence that the "integrity" of an automated system reflects the qualities given to it by the developers, much like the integrity of a human consists of their "set of values" [23]. Recall Kim et al.'s findings that while internal blame was effective at repairing trust after ability-based violations, it was not effective after integrity-based violations. Whereas admitting to an ability violation may be viewed as an isolated incident, authors propose that admitting to an integrity violation reflects on more central aspects of the trustee's character, implying that they cannot be trusted in the future [15]. While internal blame by the computer may signal an isolated, ability-based violation of trust, blame of the developers may reflect deep-seated problems with the system's integrity.

If framing errors as caused by the computer itself signals a lack of system ability, users may regard the violation as impermanent, expecting that the system's future behavior in similar circumstances will be different. This may lead to misuse [29] of the system in a situation where the error does repeat itself. Systems such as the Amazon Echo that are designed for particularly anthropomorphic interactions (e.g., reference by a human name, use of speech) may be prone to such overtrust if users tend toward perceiving them as independent actors rather than programmed machines.

On the other hand, if framing system errors as caused by the developers signals a lack of system integrity, users may regard the violation as a more permanent flaw. This may lead to disuse [29] of the system when it can actually help. Figure 3 demonstrates this initial tendency toward undertrust, where in the low reliability condition, pseudo-external participants tended to decrease Round 2 allocation (i.e., First Calibration) more than the other blame conditions. This may have been

due to reduced integrity and benevolence perceptions associated with blame of the developers, and is in line with the idea that trust in machines is initially based on faith, and later on perceptions of dependability and predictability that develop as the relationship progresses [11, 20]. Roughly equal allocation across blame conditions toward the end of the game suggests that experience with the system eventually informed participants' expectations more than perceptions based on the message.

Our study, observing an automated system with relatively minimal social cues as in previous CASA research [32], demonstrated that a conceptual separation of the system from its developers appears to play a role in users' perceptions. As systems become increasingly autonomous, understanding how users consider them as products of human design versus independent actors will only grow in importance. It is critical that trust repair mechanisms consider how this separation impacts calibration of trust.

CONCLUSION

In this study, we sought to observe whether users distinguish between an automated system and its developers when evaluating trust. We designed a game in which participants collaborated with automation in an image classification task. A high or low reliability system attributed blame for its errors internally, pseudo-externally, or externally. As expected, we found a main effect of reliability on both behavioral trust and trusting perceptions. Moreover, we found that internal blame by the system and blame of the developers was perceived differently. These findings suggest that, when it comes to trust, the apple *does* fall far from the tree—automated systems are not treated merely as reflections of their developers, but as distinct social actors. This notion is critical for designers to ensure that users are able to accurately gauge the trustworthiness of systems, and for fostering a future of healthy human-machine relationships.

Acknowledgments

This material is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-15-1-0490.

APPENDIX

Survey Items

Perceived Trustworthiness Characteristics (Trusting Beliefs)
Adapted from [22, 24]. Rated on a 7-point Likert scale from "Strongly Disagree" to "Strongly Agree."

Ability

- The system is competent and effective in identifying vehicles.
- The system performs its role of identifying vehicles very well.
- Overall, the system is a capable and proficient means for identifying vehicles.
- In general, the system is very knowledgeable about identifying vehicles.

Integrity

- The system is truthful in its dealings with me.
- I would characterize the system as honest.
- The system keeps its commitments.
- The system is sincere and genuine.

Benevolence

- I believe that the system acts in my best interest.
- When I require help, the system does its best to help me.
- The system is concerned about my well-being, not just its own.

Trust in Automation
Adapted from [13]. Rated on a 7-point scale from "Not At All" to "Extremely" with the prompt "Please rate intensity of your feeling of trust, or your impression of the system while operating it."

- The system is deceptive
- The system behaves in an underhanded (concealed) manner
- I am suspicious of the system's intent, action, or outputs
- I am wary of the system
- The system's actions will have a harmful or injurious outcome
- I am confident in the system
- The system provides security
- The system has integrity
- The system is dependable
- The system is reliable
- I can trust the system
- I am familiar with the system

REFERENCES

[1] Bernard Barber. 1983. The logic and limits of trust. (1983).

[2] Izak Benbasat and Weiquan Wang. 2005. Trust in and adoption of online recommendation agents. *Journal of the association for information systems* 6, 3 (2005), 4.

[3] Victoria Buchholz, Philipp Kulms, and Stefan Kopp. 2017. It's (Not) Your Fault! Blame and Trust Repair in Human-Agent Cooperation. (2017).

[4] Ross Buck. 2014. *Emotion: A biosocial synthesis*. Cambridge University Press.

[5] Shih-Yi Chien, Zhaleh Semnani-Azad, Michael Lewis, and Katia Sycara. 2014. Towards the development of an inter-cultural scale to measure trust in automation. In *International Conference on Cross-Cultural Design*. Springer, 35–46.

[6] Jacob Cohen. 1973. Eta-squared and partial eta-squared in fixed factor ANOVA designs. *Educational and psychological measurement* 33, 1 (1973), 107–112.

[7] Ewart J de Visser, Richard Pak, and Tyler H Shaw. 2018. From 'automation' to 'autonomy': the importance of trust repair in human–machine interaction. *Ergonomics* (2018), 1–19.

[8] Batya Friedman. 1995. "It's the computer's fault": reasoning about computers as moral agents. In *Conference companion on Human factors in computing systems*. ACM, 226–227.

[9] Joseph A Gliem and Rosemary R Gliem. 2003. Calculating, interpreting, and reporting Cronbach's alpha reliability coefficient for Likert-type scales. Midwest Research-to-Practice Conference in Adult, Continuing, and Community.

[10] Victoria Groom, Jimmy Chen, Theresa Johnson, F Arda Kara, and Clifford Nass. 2010. Critic, compatriot, or chump?: Responses to robot blame attribution. In *Proceedings of the 5th ACM/IEEE international conference on Human-robot interaction*. IEEE Press, 211–218.

[11] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors* 57, 3 (2015), 407–434.

[12] Theodore Jensen, Yusuf Albayram, Mohammad Maifi Hasan Khan, Ross Buck, Emil Coman, and Md Abdullah Al Fahim. 2018. Initial Trustworthiness Perceptions of a Drone System based on Performance and Process Information. In *Proceedings of the 6th International Conference on Human-Agent Interaction*. ACM, 229–237.

[13] Jiun-Yin Jian, Ann M Bisantz, and Colin G Drury. 2000. Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics* 4, 1 (2000), 53–71.

[14] Ing-Marie Jonsson, Clifford Nass, Jack Endo, Ben Reaves, Helen Harris, Janice Le Ta, Nicholas Chan, and Sean Knapp. 2004. Don't blame me I am only the driver: impact of blame attribution on attitudes and attention to driving task. In *CHI'04 extended abstracts on Human factors in computing systems*. ACM, 1219–1222.

[15] Peter H Kim, Donald L Ferrin, Cecily D Cooper, and Kurt T Dirks. 2004. Removing the shadow of suspicion: the effects of apology versus denial for repairing competence-versus integrity-based trust violations. *Journal of applied psychology* 89, 1 (2004), 104.

[16] John Lee and Neville Moray. 1992. Trust, control strategies and allocation of function in human-machine systems. *Ergonomics* 35, 10 (1992), 1243–1270.

[17] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[18] Roy J Lewicki and Chad Brinsfield. 2017. Trust repair. *Annual Review of Organizational Psychology and Organizational Behavior* 4 (2017), 287–313.

[19] Xin Li, Traci J Hess, and Joseph S Valacich. 2008. Why do we trust new technology? A study of initial trust formation with organizational information systems. *The Journal of Strategic Information Systems* 17, 1 (2008), 39–71.

[20] Poornima Madhavan and Douglas A Wiegmann. 2007. Similarities and differences between human–human and human–automation trust: an integrative review. *Theoretical Issues in Ergonomics Science* 8, 4 (2007), 277–301.

[21] Maria Madsen and Shirley Gregor. 2000. Measuring human-computer trust. In *11th australasian conference on information systems*, Vol. 53. Citeseer, 6–8.

[22] Roger C Mayer and James H Davis. 1999. The effect of the performance appraisal system on trust for management: A field quasi-experiment. *Journal of applied psychology* 84, 1 (1999), 123.

[23] Roger C Mayer, James H Davis, and F David Schoorman. 1995. An integrative model of organizational trust. *Academy of management review* 20, 3 (1995), 709–734.

[24] D Harrison McKnight, Vivek Choudhury, and Charles Kacmar. 2002. Developing and validating trust measures for e-commerce: An integrative typology. *Information systems research* 13, 3 (2002), 334–359.

[25] Tim R Merritt, Kian Boon Tan, Christopher Ong, Aswin Thomas, Teong Leong Chuah, and Kevin McGee. 2011. Are artificial team-mates scapegoats in computer games. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*. ACM, 685–688.

[26] Youngme Moon and Clifford Nass. 1998. Are computers scapegoats? Attributions of responsibility in human-computer interaction. *International Journal of Human-Computer Studies* 49, 1 (1998), 79–94.

[27] Clifford Nass and Youngme Moon. 2000. Machines and mindlessness: Social responses to computers. *Journal of social issues* 56, 1 (2000), 81–103.

[28] Clifford Nass, Jonathan Steuer, and Ellen R Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 72–78.

[29] Raja Parasuraman and Victor Riley. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human factors* 39, 2 (1997), 230–253.

[30] S Joon Park, Craig M MacDonald, and Michael Khoo. 2012. Do you care if a computer says sorry?: user experience design through affective messages. In *Proceedings of the designing interactive systems conference*. ACM, 731–740.

[31] Daniel B Quinn, Richard Pak, and Ewart J de Visser. 2017. Testing the Efficacy of Human-Human Trust Repair Strategies with Machines. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 61. SAGE Publications Sage CA: Los Angeles, CA, 1794–1798.

[32] Byron Reeves and Clifford Ivar Nass. 1996. *The media equation: How people treat computers, television, and new media like real people and places.* Cambridge university press.

[33] Paul Robinette, Ayanna M Howard, and Alan R Wagner. 2015. Timing is key for robot trust repair. In *International Conference on Social Robotics*. Springer, 574–583.

[34] Julian B Rotter. 1980. Interpersonal trust, trustworthiness, and gullibility. *American psychologist* 35, 1 (1980), 1.

[35] Kelly Satterfield, Carryl Baldwin, Ewart de Visser, and Tyler Shaw. 2017. The Influence of Risky Conditions in Trust in Autonomous Systems. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 61. SAGE Publications Sage CA: Los Angeles, CA, 324–328.

[36] Alexander Serenko. 2007. Are interface agents scapegoats? Attributions of responsibility in human–agent interaction. *Interacting with computers* 19, 2 (2007), 293–303.

[37] S Shyam Sundar and Clifford Nass. 2000. Source orientation in human-computer interaction: Programmer, networker, or independent social actor. *Communication research* 27, 6 (2000), 683–703.

[38] Edward C Tomlinson and Roger C Mayer. 2009. The role of causal attribution dimensions in trust repair. *Academy of Management Review* 34, 1 (2009), 85–104.

[39] Jeng-Yi Tzeng. 2004. Toward a more civilized design: studying the effects of computers that apologize. *International Journal of Human-Computer Studies* 61, 3 (2004), 319–345.

[40] Bernard Weiner. 1985. An attributional theory of achievement motivation and emotion. *Psychological review* 92, 4 (1985), 548.

[41] Christopher D Wickens and Stephen R Dixon. 2007. The benefits of imperfect diagnostic automation: A synthesis of the literature. *Theoretical Issues in Ergonomics Science* 8, 3 (2007), 201–212.